Politecnico di Torino

Laurea Magistrale in

Ingegneria del cinema e dei mezzi di comunicazione

Tesi di Laurea Magistrale

AI-DRIVEN VIDEO GENERATION: ricerca, analisi e utilizzo dei tool di AI video generativa che stanno rivoluzionando il mondo della comunicazione

Relatori

Prof. Andrea BOTTINO

Dott. Francesco STRADA

Candidato

Stefano TAGLIENTE

Aprile 2024

Abstract

Il lavoro di tesi mira ad offrire una ricerca, un'analisi e un approfondimento sul tema dell'intelligenza artificiale e, in particolare, su una sotto-tipologia che sta assumendo sempre più importanza in ambito tecnologico e creativo, l'AI generativa.

Il lavoro è stato svolto in Reply, azienda italiana di consulenza informatica e tecnologica, e comprende un'introduzione generale sull'AI e sulle sue diverse sfaccettature. In seguito, è stata condotta una ricerca sulla logica e sul funzionamento che riguarda l'AI generativa. Quindi, sono stati individuati i diversi approcci con cui la Generative AI si presenta, dalla teoria dei GAN a quella dei modelli di diffusione, fino ad esplorare anche le recenti scoperte e ricerche elaborate da team aziendali. Ne è un esempio il metodo Emu Video ed è presente una sezione di approfondimento sul funzionamento di tool come Runway.

Terminata la prima fase di lavoro di tesi, si è passati ad effettuare una ricerca e un'analisi sui diversi tool dell'AI generativa come: Runway, Kaiber, Genmo, Pika.

In particolare, sono state effettuate diverse prove e generazioni che potessero dare un contributo sia dal punto di vista quantitativo che qualitativo. Quantitativo perché si sono ottenuti diversi risultati in output, prima di andare ad individuare i risultati migliori in base agli input dati in ingresso. Qualitativo, invece, dal momento che sono stati selezionati solo alcuni video come ben realizzati e, su questi, sono state date delle considerazioni sia soggettive che oggettive.

Nello specifico, i video ottenuti in output dai vari tool AI generatori, vengono confrontati tra di loro per capire quale tool è stato più efficiente in termini di qualità risolutiva e di fedeltà relativa agli elementi inseriti in input.

Infatti, le generazioni effettuate sono di tipo Text-to-Video Generation e questo significa che i vari video di output sono stati generati a partire da un prompt testuale inserito in input.

Tra i parametri presi in considerazione per la valutazione soggettiva dei diversi tool, troviamo: qualità del rendering e della risoluzione; gli effetti di luci e ombre; il 3D; il fotorealismo; la gamma di colori; la precisione dei dettagli e delle textures degli oggetti generati.

Per le valutazioni oggettive sono state poste delle domande a un numero di valutatori umani.

In seguito, una volta condotti i test e capiti quali applicativi fossero i migliori, è iniziata la fase di produzione di video AI all'interno dell'azienda Reply, in cui il candidato ha intrapreso uno stage collaborativo. Nello specifico, a partire da alcuni progetti svolti in Reply, è stata prodotta una serie video originale, interamente generata in AI: script, storyboard, image e video generation, voice e audio generation, video editing indicano il flusso di lavoro seguito per la realizzazione.

Dai video AI realizzati, prima nei test e poi per l'azienda, si capisce come la nuova tecnologia emergente della Generative AI sta cambiando il modo di realizzare i contenuti multimediali e, soprattutto, sta rivoluzionando il mondo della comunicazione.

Il futuro si prepara ad essere rappresentato da questo nuovo medium creativo e questo lavoro di tesi mira ad essere una base approfondita per uno scenario che sarà presto modificato dalle nuove tecnologie emergenti in ambito AI generativa.

Indice

1. Introduzione

1.1 Scenario tecnologico

Nel cuore della rivoluzione digitale, la convergenza tra l'Intelligenza Artificiale (*Artificial Intelligence nella traduzione inglese*) e la produzione di contenuti multimediali ha delineato un nuovo orizzonte nell'evoluzione tecnologica. Laddove un tempo la creazione di video e contenuti visivi richiedeva competenze umane approfondite e sforzi considerevoli, l'emergere dell'AI generativa ha trasformato radicalmente questa narrativa.

Nelle fasi iniziali dello sviluppo tecnologico, l'elaborazione e la produzione multimediale erano vincolate da risorse computazionali limitate e algoritmi rigidi e non molto veloci. La creazione di contenuti visivi richiedeva la programmazione dettagliata di ogni aspetto, costringendo gli sviluppatori e gli addetti al lavoro multimediale ad affrontare una manualità pratica che risultava spesso limitante in termini di creatività e originalità nella realizzazione e produzione di contenuti multimediali.

L'introduzione dell'Intelligenza Artificiale ha segnato una svolta significativa, consentendo alle macchine di apprendere da dati esistenti e di ottimizzare il processo creativo. Nel contesto della produzione di contenuti multimediali, quali video con immagini e audio, l'uso dell'AI ha significato una maggiore automazione nei processi di editing, un miglioramento delle prestazioni durante la realizzazione dei contenuti, il riconoscimento delle immagini e dei testi trattati e ha garantito una maggiore comprensione del contesto visivo che si vuole rappresentare.

L'ascesa dell'AI generativa ha rappresentato il culmine di questa evoluzione tecnologica e creativa, portando con sé la capacità di non solo comprendere e riprodurre immagini, testi e video esistenti ma anche di creare contenuti visivi completamente nuovi. Attraverso algoritmi di diverso tipo, l'AI generativa può ora concepire e produrre diversi contenuti multimediali in modo autentico e innovativo.

Nel contesto attuale, specifico e avanzato, della generazione di video, le recenti innovazioni nell'AI generativa aprono le porte a scenari inimmaginabili. Dalla simulazione e realizzazione di ambienti virtuali, all'animazione di personaggi ed oggetti ricreati digitalmente, l'AI generativa si erge come una forza creativa, plasmando il futuro della produzione video in modi che solo pochi anni fa sembravano fantascienza.

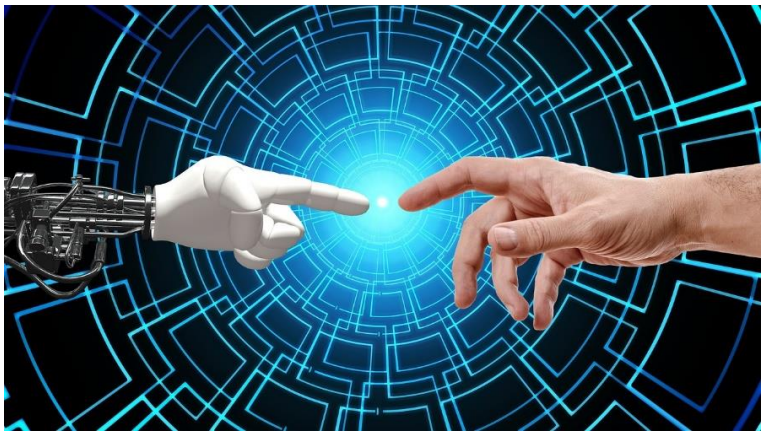

Figura 1.1 Testimonianza dell'unione tra AI e uomo per un futuro da condurre insieme

1.2 Panoramica della tesi

In questa tesi, esploreremo approfonditamente come l'AI generativa stia rivoluzionando il mondo della comunicazione e il modo di approcciarsi alla creazione e realizzazione di video innovativi. Ricercando, analizzando e utilizzando diversi strumenti o tool AI di video, image, text e audio generation, si sono comprese le caratteristiche e le funzionalità che sono offerte dall'AI generativa ma anche le logiche e gli algoritmi che governano e regolano questo nuovo e ampio medium creativo.

Tramite l'utilizzo di alcuni principali tool di Intelligenza Artificiale sviluppati e rilasciati fino a questo momento, l'AI generativa è stata approcciata e testata nelle diverse aree di intervento: dalla generazione di testi e immagini, alla generazione di video e audio con la presenza di operazioni di sintesi vocale precise e accurate.

Attraverso la fruizione degli strumenti generativi è stato possibile generare dei contenuti efficienti e di qualità, con un flusso di lavoro particolare svolto per ottenere dei video AI.

Alcuni video AI realizzati sono stati considerati per svolgere delle fasi di test e confronto, utili a valutare, con metodologie differenti, le effettive e diverse generazioni ottenute.

Inoltre, diversi video AI sono stati prodotti anche per rappresentare la fase di produzione e realizzazione di una serie video originale interamente sviluppata con l'ausilio dell'AI.

Al lavoro di tesi puramente pratico, rappresentato dalla realizzazione di video AI elaborati e completi, si affianca un lavoro di ricerca e analisi dei diversi tool AI trattati per parametri e si approfondisce la comprensione degli approcci, con le relative logiche, adottati dall'AI generativa per generare i contenuti.

1.3 Obiettivi della tesi

Gli obiettivi che si vogliono raggiungere con il seguente lavoro di tesi, rappresentato dalla ricerca, dall'analisi e dagli utilizzi di differenti tool AI video generativi trattati, riguardano:

- La comprensione e l'applicazione di alcune possibili metodologie generative che si possono attuare per realizzare dei contenuti digitali tramite l'AI generativa;
- Le scelte oculate di quali strumenti prendere in considerazione per le generazioni guidate dall'AI generativa;
- La risoluzione e l'approccio alle sfide e alle opportunità creative che l'AI video generativa propone;
- La conoscenza degli approcci, delle logiche e degli algoritmi che guidano l'AI generativa.

L'elaborato vuole rappresentare le basi per una prima versione approfondita e dettagliata delle diverse possibilità creative, procedurali e innovative che l'AI generativa ha da offrire, con una focalizzazione dettagliata relativa a contenuti come: testo, immagini, video e audio.

2. Stato dell’arte

2.1 La nuova tecnologia: l’AI

L’AI, acronimo di Artificial Intelligence, è una delle più recenti scoperte diffuse nell’ambito tecnologico e, ormai da tempo, nel quotidiano. L’uso dell’AI trova applicazione in diversi contesti aziendali: dall’informatica alla finanza, dal marketing all’arte generativa e multimediale, da modelli di business all’ambito medico e sanitario.
Nell'epoca moderna, l'Intelligenza Artificiale si distingue come una pietra miliare tecnologica che ha trasformato radicalmente il modo in cui affrontiamo problemi complessi e prendiamo decisioni. Basata sull'idea di sviluppare sistemi in grado di apprendere dai dati e migliorare autonomamente, l'AI ha conquistato una presenza onnipresente in diversi settori. Attraverso algoritmi avanzati, reti neurali profonde e tecniche di apprendimento automatico, l'AI ha dimostrato capacità straordinarie nella comprensione del linguaggio testuale e verbale, nella visione artificiale di contenuti come immagini e video e nell'elaborazione dei dati. Questa disciplina in continua evoluzione offre non solo soluzioni pratiche per problemi aziendali e scientifici, ma anche un terreno fertile per l'emergere di nuovi paradigmi di interazione tra uomini e macchine, aprendo la strada a scenari futuristici che solo pochi decenni fa sembravano appartenere a un altro mondo.
La ricerca attuale nell'ambito dell'AI si concentra su sfide stimolanti come l'interpretabilità dei modelli di diffusione, l'etica nell'uso dell'AI e la creazione di sistemi sempre più autonomi e adattabili a diversi scenari d’uso.

2.1.1 Funzionamento generale dell’AI

L’AI svolge le operazioni di generazione tramite dei modelli di apprendimento che elaborano e creano le informazioni a partire da dati esistenti e, una volta che hanno compreso quello che devono generare tramite queste conoscenze di contenuti presenti in specifici dataset, tendono a generare risultati in output nuovi e unici.

Il funzionamento dell’intelligenza artificiale viene definito tramite quattro differenti livelli funzionali, in grado di svolgere le operazioni che l’AI è chiamata ad eseguire:

- *Comprensione*: questa deriva dalla competenza di apprendere e replicare la connessione tra dati ed eventi. Attraverso questa capacità, l'AI può, ad esempio, identificare testi, immagini, video, audio e voci per elaborare informazioni specifiche in risposta a una richiesta particolare;
- *Ragionamento*: questo emerge dalla competenza dei sistemi AI di connettere logicamente e autonomamente i dati acquisiti, utilizzando una serie di algoritmi matematici attentamente configurati;
- *Apprendimento*: si manifesta attraverso sistemi AI capaci di analizzare i dati in ingresso per produrre un output accurato. Un esempio è rappresentato dai sistemi di apprendimento automatico (Machine Learning) che impiegano tecniche specifiche per acquisire conoscenze da un contesto informativo specifico, al fine di eseguire funzioni particolari;
- *Interazione*: si verifica attraverso i sistemi HMI (Human Machine Interaction), nei quali l'AI svolge un ruolo cruciale nella sua relazione con gli esseri umani. Un esempio comune di questa dinamica è rappresentato dal Natural Language Processing (NLP), un insieme di tecnologie basate sull'intelligenza artificiale che facilita l'instaurarsi di una comunicazione verbale tra l'uomo e la macchina, sfruttando il linguaggio naturale. Questo fenomeno è evidente nei casi dei chatbot più avanzati, come chatGPT di OpenAI.

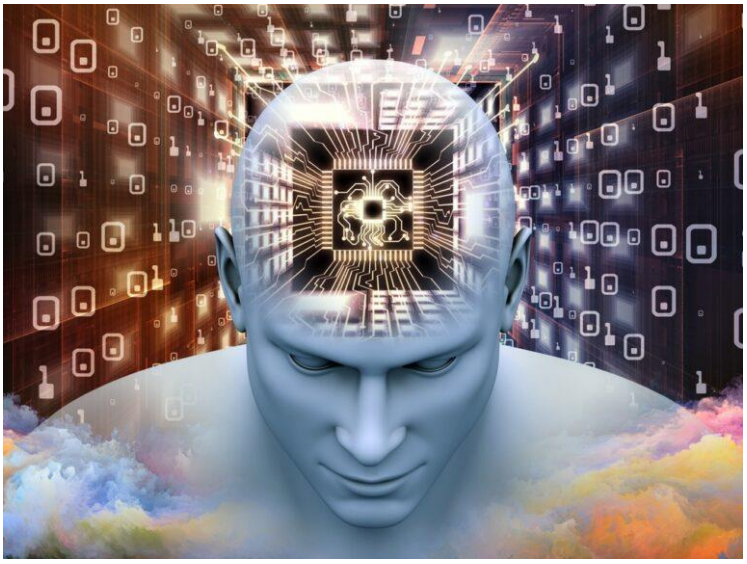

Figura 2.1 Individuo umano virtuale (AI) che elabora le diverse informazioni percepite

2.1.2 Concetti base AI: Machine Learning vs Deep Learning

Tra le diverse terminologie legate all'ambito AI, è bene distinguere i concetti di Machine Learning e di Deep Learning.

Tra le due tecniche di apprendimento vi è una sostanziale differenza e questa si manifesta anche nei modelli di apprendimento su cui queste sono basate.

Il Machine Learning (ML) è un sistema di apprendimento automatico basato sull'intelligenza artificiale. Questo sistema è in grado di raccogliere una vasta gamma di dati (gli input) per addestrare una macchina, che gradualmente acquisisce abilità crescenti nell'esecuzione di un compito specifico (gli output) senza la necessità di una programmazione preventiva. Quello che contraddistingue un sistema di Machine Learning è la sua capacità di apprendere, commettere errori e migliorarsi progressivamente da tali errori, evolvendo costantemente per diventare sempre più preciso nelle simulazioni che può produrre in modo autonomo.

Il modello di apprendimento di un sistema di ML è piuttosto vario e si basa su tre principali classi di algoritmi:

- *Con supervisione didattica*: in cui il sistema apprende mediante una correlazione tra input e output da cui impara come prendere una decisione;
- *Senza supervisione didattica*: in questo contesto, il processo di apprendimento avviene attraverso l'analisi dei risultati, senza un collegamento diretto tra input e output. Dunque, il focus è affidato esclusivamente sulla valutazione degli output, che permette di mappare i risultati delle decisioni in un determinato contesto. Questo approccio è particolarmente rilevante nei sistemi di Machine Learning chiamati a fornire soluzioni in assenza di guida diretta o supervisione didattica;
- *Con rinforzo*: il reinforcement learning rappresenta un approccio di apprendimento basato sul merito, in quanto l'AI viene premiata solo quando le sue valutazioni producono risultati in linea con le aspettative. Questo metodo permette di perfezionare l'addestramento di un sistema di Machine Learning poiché abilita l'AI a distinguere fondamentalmente tra decisioni corrette e errate, guidata dal feedback positivo o negativo ottenuto durante il processo di apprendimento.

Il Deep Learning (DL) consiste in modelli di apprendimento ispirati al funzionamento del cervello umano.
Non si tratta di un metodo di allenamento strettamente basato sulla relazione tra un input e un output, come nel caso del ML, quanto di un sistema che utilizza gli input per arrivare a emulare il comportamento del cervello umano.

Il Deep Learning si fonda sulle reti neurali profonde, che si distinguono per la presenza di numerosi strati di calcolo. Questi strati, costituiti da un numero significativo di livelli, richiedono sforzi computazionali considerevoli al fine di creare un ambiente che rifletta le complesse connessioni neurali presenti nel cervello umano, sebbene molte di queste connessioni rimangano ancora in gran parte sconosciute.

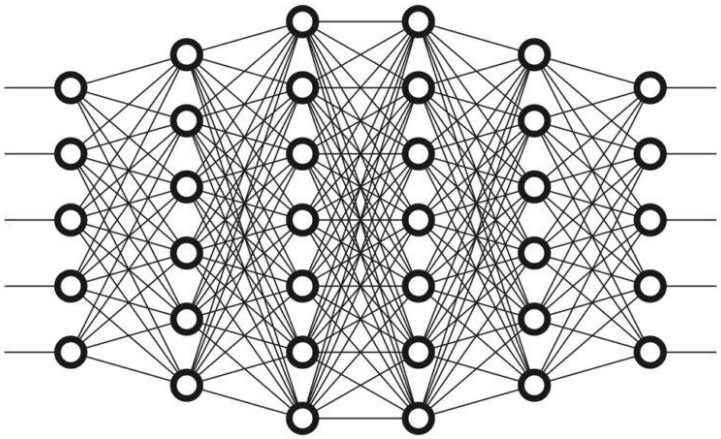

Figura 2.2 Una serie di connessioni neurali relative alle reti neurali su cui si basa il deep learning

2.2 L'AI Generativa

Tra le applicazioni emergenti dell'Intelligenza Artificiale compare l'AI Generativa (*Generative AI in inglese*). Questo nuovo medium tecnologico è trattato all'interno del lavoro di tesi.
La Generative AI è in grado di produrre dati sintetici virtuali e di supportare le capacità e le attività creative dell'essere umano.
Rientrano tra le tecnologie collegate alla Generative AI, quelle che consentono a un sistema di Machine Learning, opportunamente istruito e allenato attraverso l'utilizzo di dataset a tema, di produrre e generare contenuti artificiali di diverso tipo.
Dunque, la Generative AI è un tipo di intelligenza artificiale che può generare un'ampia varietà di dati e contenuti, come le immagini, i video, gli audio e i segnali vocali, i testi e, di recente, anche i modelli 3D, contribuendo ad essere utile nell'ambito della computer grafica, della computer animation e della produzione di video in generale.
L'AI Generativa è in grado di produrre contenuti altamente realistici, complessi e precisi che imitano, e talvolta superano, quelli che si ottengono con la creatività umana.
Questa scienza creativa non solo può potenziare le capacità analitiche e decisionali dell'uomo ma anche migliorare la creatività e la produzione di contenuti.

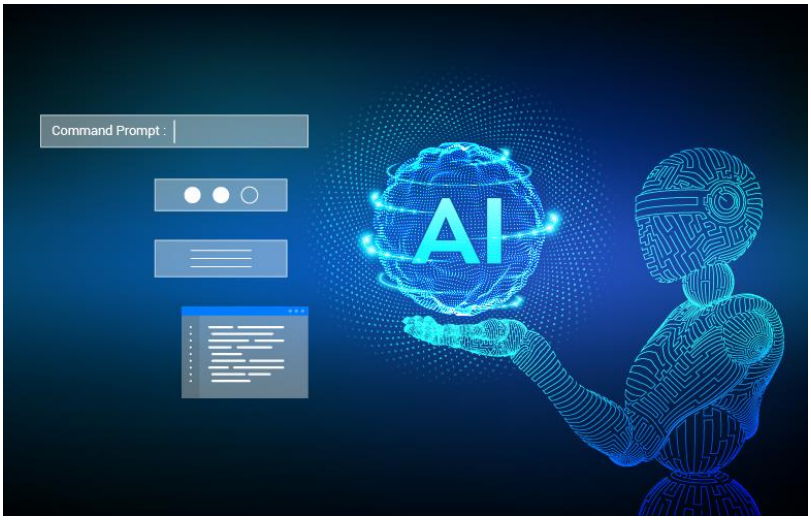

Figura 2.3 Da un lato l'area di comando con il prompt usato per la generazione di contenuti con il caricamento di questi e dall'altro la macchina AI che governa l'operazione

2.2.1 Ambiti di applicazione dell'AI Generativa

L'AI Generativa può essere uno strumento e un elemento che valorizza e facilita il lavoro alle diverse aziende e agenzie di comunicazione: realtà come quelle del gaming, dell'intrattenimento, del design, dell'automotive e dell'ambito informatico sono solo un esempio di contesti e ambiti in cui l'AI generativa viene utilizzata.

Le recenti scoperte e i continui aggiornamenti riguardanti l'AI generativa nel mondo multimediale, hanno portato allo sviluppo di nuovi strumenti o tool AI di generazione di diverse tipologie di contenuti. Le generazioni offerte dall'AI Generativa riguardano diversi contenuti come: testi, immagini, video, audio e voci sintetiche.

L'utilizzo dell'AI generativa consente diverse generazioni di contenuti che portano ad ottenere risultati differenti.

Le generazioni consentite dall'applicazione dell'AI Generativa sono:

- La generazione di testo (*Text Generation*) che prevede l'utilizzo di modelli di apprendimento automatico (ML) per generare nuovi testi basati su modelli appresi da dataset di testo esistenti. I modelli utilizzati per la generazione del testo possono essere le reti neurali ricorrenti (RNN) e, sviluppati di recente, i trasformatori, che hanno rivoluzionato il campo grazie alla loro estesa capacità di attenzione.
  La generazione del testo ha numerose applicazioni nel campo dell'elaborazione del linguaggio naturale verbale, dei chatbot e della creazione di contenuti.
  Tra i tool AI più usati in questo contesto emerge ChatGPT (Generative Pre-trained Transformer), sviluppato da OpenAI, che utilizza la Text Generation per generare risposte simili a quelli degli umani attraverso la creazione di conversazioni via chat. Recentemente consente anche di effettuare generazioni di immagini ma a partire da certe versioni del tool (ChatGPT 4).
- La generazione di immagini (*Image Generation*) che è un processo che consiste nell'utilizzo di algoritmi di apprendimento profondi (deep learning) come le GAN e il più recente Stable Diffusion, per creare nuove immagini che sono visivamente simili alle immagini del mondo reale con un alto grado di somiglianza e realisticità degli elementi. La generazione di immagini può essere usata per aumentare la quantità di contenuti artistici di diverso tipo che hanno una diversa qualità, ma anche per dare l'input alla generazione e creazione di diversi video che sono prodotti a partire dalle immagini di riferimento generate.
  Tra i tool AI più diffusi per la generazione di immagini abbiamo Midjourney, DALL-E e Leonardo.

- La generazione di video (*Video Generation*) che è una tipologia di generazione che si basa su metodi di apprendimento profondi come le GAN e Video Diffusion per generare nuovi video tramite, solitamente, la predizione di frame che si basa su frame precedenti e si fa uso della video prediction sui frame. La generazione dei video può essere usata in diversi campi, tra cui l'intrattenimento, la televisione e il cinema.
  Alcuni tool AI usati nel contesto video sono: Runway, Kaiber, Pika e Genmo.
- La generazione di voce, segnali vocali o speech, audio (*Voice e Audio Generation*) che è utilizzata da sola oppure trova applicazione insieme alla generazione dei video, in cui al video generato si inserisce una certa voce artificiale generata che si avvicina molto alla voce umana come timbro e frequenza (pitch). Questo segnale vocale generato, può essere usato in voice over come voce di sottofondo e accompagnamento al video oppure come speech e quindi tramite parlato del soggetto che appare all'interno del video AI. I modelli usati per la generazione vocale possono essere prodotti da modelli basati su trasformatori. La generazione della voce trova applicazione nella conversione text-to-speech, negli assistenti virtuali e nella clonazione di voce per generare e ottenere voci che sono il più possibile simili a quelle di una determinata persona.
  Tool come Synthesia ed Heygen sono utilizzati sia per la parte di generazione video che per quella di generazione voce con la presenza di avatar che parlano. Invece, ElevenLabs è un tool AI usato esclusivamente per la generazione di voce e segnali vocali.

Come si può intuire dalle categorie di generazioni AI, sono stati sviluppati degli strumenti generativi appropriati. Questi applicativi, come anche tanti altri che si stanno implementato e rilasciando, hanno significativamente avanzato e migliorato le loro funzionalità interne e hanno cambiato il modo di utilizzare l'AI generativa in maniera efficace ed efficiente.

Oltre a questi contesti di applicazione, la Generative AI ha un range vasto di intervento che va oltre la generazione di testo, immagini, video, voce, audio. Per esempio, può essere usata esclusivamente per la generazione di musica, per lo sviluppo di videogiochi, per l'assistenza sanitaria e altro.

Questi continui sviluppi hanno aperto a nuove possibilità di utilizzare l'AI Generativa per risolvere problemi complessi tra cui la creazione di diverse composizioni artistiche e multimediali ma anche assistere ed aiutare la conduzione della ricerca scientifica.

Figura 2.4 chatGPT text generator

Figura 2.5 Midjourney image generator

Figura 2.6 Runway vGen-2 video generator

Figura 2.7 ElevenLabs voice generator

2.2.2 L'approccio GAN

Nel campo dell'Intelligenza Artificiale generativa, una delle tecnologie più emblematiche è rappresentata dalle GAN (Generative Adversarial Networks).

Le GAN sono un tipo di architettura di rete neurale artificiale utilizzata nell'ambito dell'apprendimento automatico (Machine Learning) ed è collegato al contesto dell'AI generativa e della generazione di differenti dati sintetici virtuali. Queste tipologie di reti, che indicano una nuova tecnologia ma anche un nuovo approccio di generare contenuti con l'AI, sono state proposte nel 2014.

Le GAN sono composte da due reti neurali principali: il Generatore e il Discriminatore. Queste vengono addestrate simultaneamente attraverso un processo competitivo.

Il Generatore (*o Generator in inglese*) è una rete che cerca di generare dati, come immagini, suoni o testi, da rumore casuale. L'obiettivo del generatore è produrre dati che siano indistinguibili dai dati reali. In altre parole, il Generatore è il cuore della GAN e svolge un ruolo cruciale nella creazione di dati artificiali. La sua funzione principale è quella di tradurre un input casuale, spesso chiamato "rumore", in dati che dovrebbero assomigliare il più possibile ai dati reali presenti nel dataset di addestramento della rete.

Per comprendere meglio il funzionamento del Generatore, e di una porzione di GAN, è utile definire alcuni punti chiave:

- *Input casuale*: il Generatore inizia con un input casuale, solitamente sotto forma di vettore di numeri casuali. Questo input casuale è fondamentale per garantire la varietà nella generazione di dati;
- *Mapping a dati realistici*: il Generatore utilizza il suo modello, spesso basato su una rete neurale, per mappare l'input casuale in dati che dovrebbero sembrare autentici. Ad esempio, se il Generatore è addestrato per generare immagini di volti umani o di luoghi, il suo compito è tradurre il rumore casuale in immagini di volti credibili o luoghi verosimili alla realtà.
- *Apprendimento*: durante il processo di addestramento, il Generatore è costantemente regolato in modo che le sue produzioni siano sempre più vicine ai dati reali del dataset di addestramento. L'obiettivo è produrre dati che siano "indistinguibili" da quelli reali quando visti da un'altra rete, chiamata Discriminatore.
- *Risultati ottimali*: l'allenamento mira a rendere il Generatore così abile da generare dati che siano difficili da distinguere da quelli reali. In altri termini, l'obiettivo è che il Discriminatore (che cerca di distinguere tra dati reali e generati) sia ingannato al massimo.
  In sintesi, il Generatore è il creatore di dati nella GAN e sfrutta un input casuale per produrre uscite che dovrebbero essere virtualmente indistinguibili dai dati reali. Questa dinamica di competizione con il Discriminatore contribuisce al miglioramento costante delle abilità del Generatore nel produrre dati sempre più autentici e precisi.

Invece, il Discriminatore (*o Discriminator in inglese*) è una rete che valuta se un dato è reale (proveniente dal dataset di addestramento implementato) o è generato dal generatore. Il suo obiettivo è distinguere in modo accurato tra dati reali e generati.

Il Discriminatore agisce come un "giudice" nella dinamica delle GAN. La sua responsabilità è distinguere tra dati reali provenienti dal dataset di addestramento e dati generati dal Generatore. Il Discriminatore è anch'esso una rete neurale, ma è addestrato per essere sempre più competente nel rilevare la differenza tra il "vero" e il "falso". Durante il processo di addestramento, il Discriminatore fornisce un feedback al Generatore, indicando quanto bene o quanto male sta svolgendo il suo compito. Questo feedback è essenziale perché spinge il Generatore a migliorare costantemente nel tentativo di eludere le capacità discriminative del Discriminatore.

In breve, il Generatore e il Discriminatore lavorano in tandem in un processo iterativo, creando una sorta di "competizione" in cui il Generatore cerca di diventare sempre più abile a generare dati convincenti, mentre il Discriminatore cerca di diventare sempre più abile a riconoscere la differenza

tra dati reali e generati. Questa competizione spinge entrambe le reti a migliorare continuamente le loro prestazioni ed è una "gara" continua che porta alla creazione di dati sintetici di alta qualità.

L'approccio GAN è ampiamente utilizzato per generare dati sintetici in vari campi, come la generazione di immagini, la creazione di deepfake, la sintesi e la generazione di testo e molto altro.

Nello specifico, le reti GAN sono particolarmente conosciute per il loro impiego nella produzione di deepfake, ossia contenuti audio e video in grado di sostituire in modo estremamente accurato la controparte reale.

Esempi comuni di applicazione del deepfake tramite l'AI, e le GAN, includono la sostituzione di volti originali presenti in video con altri volti presi esternamente e, in questo modo, le animazioni facciali con i nuovi volti rimangono coerenti a quelle che si avevano con i volti originali delle figure riprese.

Altro esempio comune di applicazione del deepfake è la generazione di audio falso con voci identiche a quelle del soggetto che si desidera imitare. Questa tipologia di utilizzo del deepfake risulta essere leggermente meno frequente rispetto a quella applicata nei video.

All'interno delle reti GAN, sono presenti diversi approcci specifici per la generazione di contenuti. Infatti, si potrebbe parlare di varianti specifiche di GAN come:

- DCGAN (Deep Convolutional GAN) per immagini ad alta risoluzione;
- cGAN (Conditional GAN) per controllare la generazione in base a specifici input condizionali.

Le GAN non sono l'unica tecnologia attiva nell'ambito delle Generative AI. Tra le più diffuse ritroviamo anche i transformers (es. i sistemi GPT-3 e LaMDA), molto efficaci con il linguaggio verbale umano, e i variational auto-encoders, disponibili anche in diverse app mobile che si appoggiano su quest'ultima tecnologia AI generativa.

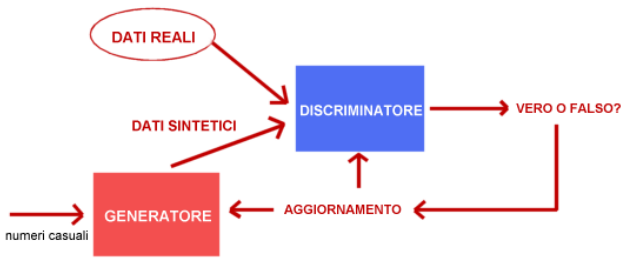

Figura 2.8 Schema di funzionamento degli elementi principali reti GAN

2.2.3 Nuovo approccio: modelli di diffusione

Nell'ambito dell'AI Generativa, oltre alla tecnologia e all'approccio delle reti GAN, un altro approccio utilizzato per la generazione di contenuti tramite l'intelligenza artificiale è quello guidato dai modelli di diffusione.

Questi modelli specifici adottati nel contesto dell'AI Generativa si riferiscono, generalmente, a dei tipi di architetture di reti neurali utilizzate per la generazione di dati sintetici.

I recenti prodotti e sistemi di AI Generativa, sviluppati da aziende come Nvidia, Google, Adobe e OpenAI, hanno considerato i modelli di diffusione come gli elementi principali utilizzati per governare la logica della generazione di contenuti. Infatti, queste aziende hanno posto i modelli di diffusione al centro dell'attenzione.

Esempi di tool che basano la loro generazione di contenuti sui modelli di diffusione sono: DALL-E di OpenAI, Stable Diffusion e Midjourney. Infatti, questi, come anche altri tool AI generatori di immagini, vengono definiti e interpretati, al tempo stesso, come degli importanti modelli di diffusione.

All'interno di questi applicativi, gli utenti forniscono un prompt testuale come input e questi modelli o tool AI generatori di immagini leggono il contenuto inserito in ingresso e lo convertono in immagini realistiche e ben definite, generando l'output di un'immagine che sia fedele il più possibile alla descrizione testuale scritta in input. Un esempio di immagine generata dall'AI, tramite una text-to-image generation, può essere quella mostrata qui di seguito:

Figura 2.9 Immagine generata con Midjourney v5

Questa immagine è stata generata in output a partire da un prompt testuale di input che cita "papaveri ben definiti e colorati delle campagne della California". La generazione dell'immagine è stata eseguita all'interno del tool AI Midjourney, che basa la generazione sull'uso dei modelli di diffusione.

2.2.4 Logica di funzionamento dei modelli di diffusione

I modelli di diffusione si basano sulla logica matematica dei principi gaussiani, sulla varianza, sulle equazioni differenziali e sulle sequenze generative da seguire per ottenere l'output finale.

Dando una definizione di modello di diffusione probabilistico, definito brevemente come modello di diffusione, si evince che esso è una catena di Markov parametrizzata addestrata per utilizzare l'inferenza variazionale per produrre campioni che corrispondono ai dati sintetici da far visualizzare dopo un tempo finito. Le transizioni o i passaggi di questa catena vengono compresi per invertire un processo di diffusione, che è una catena di Markov che aggiunge gradualmente del rumore ai dati nella direzione opposta al campionamento, fino a quando il segnale (l'immagine originale) non viene distrutto.

Quando la diffusione consiste nella presenza di piccole quantità di rumore gaussiano, è sufficiente impostare le transizioni della catena di campionamento su condizioni gaussiane, consentendo una parametrizzazione della rete neurale particolarmente semplice che tende a rimuovere progressivamente il rumore.

In poche parole, i modelli di diffusione possono generare dati simili a quelli su cui vengono addestrati, dopo aver fatto l'operazione di aggiunta di rumore (noise) e di rimozione graduale di questo (denoise). Per esempio, se il modello si allena su immagini di certi elementi, allora può generare immagini simili realistiche di questi elementi definiti nel prompt testuale.

I modelli di diffusione si ispirano al principio di funzionamento e al fondamento matematico di un modello probabilistico in grado di analizzare e prevedere il comportamento e l'andamento di un sistema che varia nel tempo.

La definizione afferma che si tratta di catene di Markov parametrizzate, istruite attraverso inferenza variazionale. Dunque, le catene di Markov rappresentano modelli matematici che descrivono un sistema in transizione da uno stato all'altro nel tempo. Il presente stato del sistema può influenzare solo la probabilità di transizione verso uno stato specifico. In altre parole, il corrente stato di un sistema contiene le potenziali transizioni che il sistema può compiere o le condizioni che può assumere in un determinato istante.

L'allenamento del modello tramite inferenza variazionale implica elaborati calcoli relativi alle distribuzioni di probabilità. L'obiettivo è identificare con precisione i parametri della catena di Markov che meglio si adattano ai dati osservati a un dato istante. Questo processo tende a minimizzare il valore della funzione di perdita del modello, rappresentante la discrepanza ("gap") tra lo stato previsto (sconosciuto) e quello osservato (noto).

Dopo l'addestramento, il modello è in grado di generare campioni che riflettono i dati osservati. Questi campioni rappresentano possibili percorsi o scenari che il sistema potrebbe seguire nel tempo, e ciascun percorso ha una probabilità associata di manifestarsi. Di conseguenza, il modello può anticipare il comportamento futuro del sistema creando una serie di campioni e determinando le probabilità corrispondenti a ciascun evento (la probabilità che tali eventi si verifichino e che utilizzino i campioni generati).

Nell'AI Generativa, i modelli di diffusione vengono interpretati come modelli generativi profondi (di deep learning) che funzionano aggiungendo del rumore (rumore gaussiano) ai dati di addestramento disponibili (noto anche come processo di diffusione in avanti) e, in seguito, inverte il processo (noto come denoising o processo di diffusione inversa) per recuperare i dati e generare l'immagine senza rumore. Il modello impara gradualmente a rimuovere il rumore. Questo processo di denoising appreso genera nuove immagini di alta qualità e risoluzione a partire da immagini semi casuali (immagini rumorose casuali).

L'interpretazione e il funzionamento dei modelli di diffusione nell'AI generativa possono essere rappresentati dalla seguente figura:

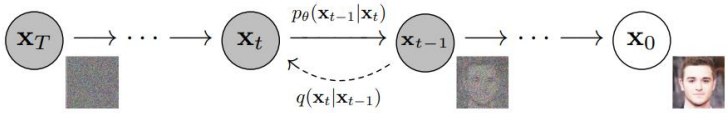

Figura 2.10 Processo di diffusione inversa con il denoising effettuato per recuperare l'immagine originale (o generare le sue variazioni) attraverso un modello di diffusione addestrato

2.2.5 Tipologie e vantaggi dei modelli di diffusione

I modelli di diffusione possono essere suddivisi in 3 categorie che contengono dei relativi modelli matematici fondamentali che sono alla base della logica dei diffusion models. Tutti e tre i modelli matematici lavorano sugli stessi principi logici fondamentali che consistono nell'aggiunta di rumore e, successivamente, nella rimozione di questo per generare nuovi campioni di dati che rappresentano le nuove immagini (utili eventualmente per i video) o i nuovi campioni audio.

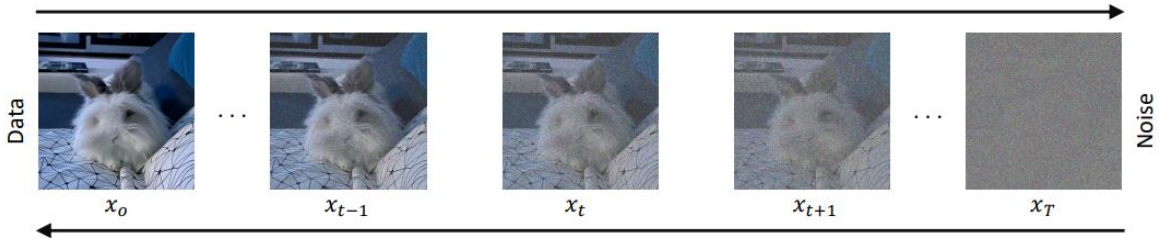

Figura 2.11 Modello di diffusione che aggiunge e rimuove rumore da un'immagine di input

Queste tre tipologie dei modelli di diffusione sono:

- Modelli probabilistici di diffusione del rumore (*DDPM acronimo inglese*): sono modelli generativi usati principalmente per rimuovere il rumore dai dati visivi come immagini o dagli audio. Nelle ricerche dei team aziendali, hanno mostrato risultati molto incoraggianti su varie attività di riduzione del rumore da immagini e audio. Ad esempio, nel cinema si stanno utilizzando moderni strumenti di elaborazione di immagini e video per migliorare la qualità della produzione adottando questi modelli.
- Modelli generativi basati su punteggio condizionato dal rumore (*SGM acronimo inglese*): questi modelli possono generare nuovi campioni da una data distribuzione. Funzionano istruendo una funzione del punteggio di stima che può stimare la densità logaritmica della distribuzione del target di riferimento (immagine o audio in input che vengono elaborati). La stima della densità presuppone che i punti dati disponibili facciano parte di un dataset sconosciuto. Questa funzione di punteggio può generare nuovi punti di stima dati dalla distribuzione iniziale dell'input. Un'applicazione degli SGM è utile nella produzione di video e audio deepfake e si dimostrano un'alternativa alle GAN usate in questo contesto. Gli SGM dimostrano capacità simili, e a volte migliori, nel generare volti o audio di alta qualità che sostituiscono quelli delle celebrità.
- Equazioni differenziali stocastiche (*SDE acronimo inglese*): essi descrivono i cambiamenti nei processi casuali relativi al tempo. Questi modelli sono usati soprattutto in fisica e nei mercati finanziari che coinvolgono fattori casuali che hanno un impatto significativo sui risultati del mercato. Per esempio, i prezzi delle materie prime sono altamente dinamici e influenzati da una serie di fattori casuali. Le SDE calcolano i derivati finanziari come le informazioni sul petrolio grezzo. Inoltre, possono modellare le fluttuazioni e calcolare in modo preciso i prezzi.

Tra i vantaggi dei modelli di diffusione, che contribuiscono all'efficacia e alla resa della generazione di contenuti, abbiamo:

- La generazione di campioni realistici in quanto i modelli di diffusione sono noti per la loro capacità di generare campioni sintetici che sono molto realistici e difficili da distinguere dai dati reali (concetto relativo anche alle GAN). Questo è particolarmente utile in applicazioni come la generazione di immagini, audio o testo.
- Apprendimento di distribuzioni complesse in quanto i modelli sono in grado di catturare distribuzioni di dati complesse. Questo significa che possono gestire dati con variazioni sottili e complessità intrinseche, rendendoli adatti a una grande varietà di applicazioni.
- Flessibilità nell'apprendimento di dati non strutturati in quanto i modelli di diffusione sono in grado di apprendere da dati non strutturati, come immagini o testi, senza la necessità di specifiche regole o modelli predefiniti. Questa flessibilità li rende adatti a contesti in cui la struttura dei dati può essere complessa o variabile.
- Addestramento non supervisionato in quanto molti modelli di diffusione possono essere addestrati in modo non supervisionato, il che significa che possono apprendere delle caratteristiche intrinseche dei dati senza etichette o guida specifica. Questo semplifica il processo di addestramento, specialmente quando i dati etichettati sono limitati o costosi da ottenere.

- Trasferimento di stile e creatività in quanto i modelli di diffusione possono essere utilizzati per trasferire stili da un tipo di dati ad un altro. Ad esempio, possono essere addestrati per applicare lo stile di un pittore a un'immagine fotografica, dimostrando una notevole creatività nell'elaborazione di contenuti.
- Applicazioni multimediali e multisensoriali in quanto questi modelli possono trattare dati di diversi tipi (immagini, audio, testo) e sono ideali per applicazioni multimediali e multisensoriali, consentendo la generazione di contenuti complessi e ricchi.

3. AI Video Generation

3.1 Tipologie AI Video Generation

All'interno dello studio di tesi e del relativo progetto condotto, l'utilizzo prevalente di AI Generativa si è basato soprattutto sulla generazione di video e sulla fruizione di tool AI video generatori (*AI video generator in inglese*).
L'AI Generativa offre la possibilità di generare video di alta qualità e di diverso tipo. I contenuti che si possono creare con questa nuova tecnologia appartengono a differenti categorie come: live-action, animazione 3D o computer animation, animazione 2D, animazione vettoriale (disegni vettoriali animati), motion graphics, contenuti cinematici di differente stile ed eventualmente altre.
Tutte queste tipologie di contenuti AI sono generate dagli algoritmi e dagli approcci che la Generative AI adotta per la generazione dei frame video, che uniti formano le relative scene video.

L'AI Generativa propone diverse tipologie di generazione video e, dunque, differenti modi di ottenere il risultato finale in output.

Le tipologie di AI Video Generation sono:

- *Text-to-Video generation*: si inserisce del testo (o prompt testuale) in input che consiste in una descrizione testuale che indica il contenuto che si desidera generare a video con l'AI, per ottenere, dunque, l'output finale. Questo approccio, assieme alla tipologia descritta immediatamente dopo, è il più utilizzato nel contesto della generazione video eseguita attraverso l'AI.
  Il motivo è relativo al fatto che l'utente, con la sua immaginazione e le sue idee espresse tramite testo scritto, può chiedere all'intelligenza artificiale generativa qualunque tipo di contenuto da generare a video e, inoltre, ha la possibilità di inserirlo in un contesto di narrazione ben specifico, in quanto la frase può essere scritta ed elaborata a dovere e in modo preciso. Quindi, la descrizione testuale offre maggiore libertà e scelta all'utente per ottenere quello che si desidera nella generazione video, e da questo punto di vista l'approccio T2V (text-to-video) generation può essere più vantaggioso rispetto alle due modalità descritte precedentemente e, inoltre, per esempio non propone il vincolo di inserire l'immagine di riferimento che definisce lo stile e il formato che il video di output assumerà a priori.
- *Image-to-Video generation*: si utilizza un'immagine di riferimento che funge da input, ma anche da frame iniziale del video che si vuole creare e generare, e su di essa vengono applicate le conseguenti animazioni per ottenere il video in output a partire dall'immagine data in ingresso. Il video, dunque, consiste nell'ottenere e nel mostrare l'immagine animata in tutte le sue parti.
- *Text+Image-to-Video generation*: si utilizza la combinazione di descrizione testuale (prompt testuale) + immagine in cui si comanda all'AI di generare un video a partire dall'immagine inserita in input e, ad essa, si aggiungono dei dettagli testuali, relativi al testo digitato, che fungono da informazioni aggiuntive all'immagine data in ingresso e servono per ottenere un video migliore e completo dal punto di vista del contenuto, delle animazioni e della narrazione delle scene.
- *Video-to-Video generation*: si utilizza un video inserito in input che consente di generare un altro video in output. Questo approccio non è offerto da tutti i tool AI di video generation e, a volte, si dimostra essere un modo non troppo affidabile di ottenere dei risultati in output, in quanto bisogna avere in ingresso un video piuttosto decente, in termini di qualità e animazioni, per poter avere in uscita un risultato ottimale che soddisfi le richieste dell'utente. In questi casi, la generazione del video in output dipende molto dal video inserito in input.
  Infatti, nel video che si genera in output, parametri come il formato e lo stile rimangono gli stessi del video inserito in input dall'utente. Questo si verifica nella maggior parte dei casi e nella buona maggioranza dei tool AI di video generation sviluppati e rilasciati. Dunque, in poche occasioni capita che il formato e lo stile possono essere variati nel video di output. Molto spesso, oltre al

formato video e allo stile visivo, anche la risoluzione rimane la stessa nella generazione video ottenuta in output a partire dal video dato in ingresso.

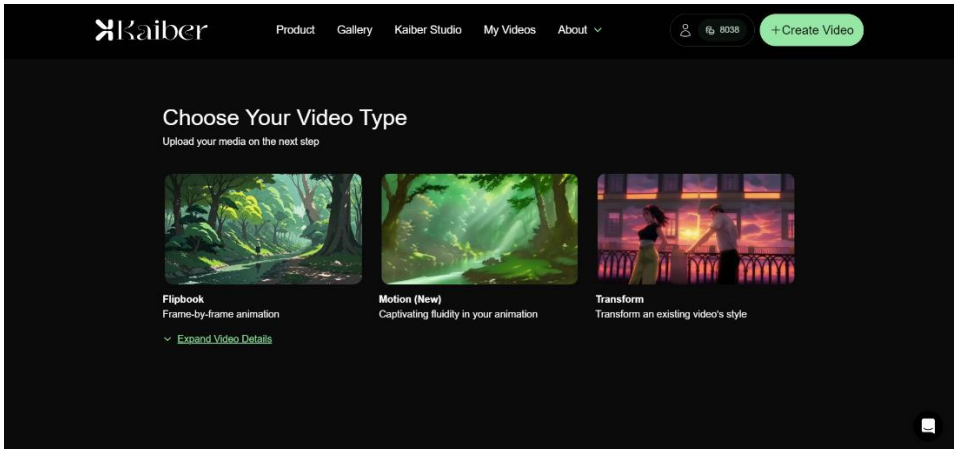

Figura 3.1 Kaiber: uno dei tool AI di video generation che consente di creare video tramite tutte e 4 le tipologie di generazione descritte

3.2 Vantaggi e svantaggi AI Video Generation

Indipendentemente dalla modalità di AI Video Generation che viene attuata per ottenere in output dei video realizzati tramite l'AI Generativa, si possono citare una serie di vantaggi e svantaggi che si possono notare quando si generano video con l'AI.

I vantaggi che si possono avere quando si applica l'AI Video Generation sono:

- La *velocità* e *l'efficienza* di produrre dei contenuti accurati. Questo è il vantaggio principale offerto dall'AI Video Generation. Infatti, i sistemi AI possono processare una grande quantità di materiale video in modo rapido ed efficiente.
  Gli algoritmi basati sul machine learning rendono la fase di processamento ed elaborazione del video automatica, facendo risparmiare del tempo sostanziale agli addetti umani che si occupano di video e immagini. Questo modo di operare naturalmente porta a una produttività incrementata e a una maggiore accelerazione nella fase di produzione e realizzazione di video.
- La *qualità* e la *risoluzione migliorata* che si possono ottenere in modo più facile e diretto. Infatti, grazie agli algoritmi avanzati, i tool AI possono automaticamente migliorare la qualità dei video elaborati singolarmente. Inoltre, l'AI Generativa può rimuovere del rumore indesiderato dai video, può stabilizzare il girato che in partenza risulta essere traballante e poco stabile, può aggiustare la fotografia con le luci e le ombre inappropriate e il contrasto dei colori presenti all'interno dell'immagine o del frame video e, infine, consente di apportare altri miglioramenti che aumentano la qualità o la risoluzione visiva complessiva dei video.
- *L'automazione dei task* nella generazione di video e immagini, ovvero si cerca di rendere automatica una certa routine di task ripetitivi, come il riconoscimento automatico, l'etichettatura e l'identificazione degli oggetti, la creazione di sottotitoli e didascalie. Questa capacità offerta dall'AI fa risparmiare del tempo considerevole per i lavoratori umani in fase di post-produzione.
- *L'abilità di generare nuovi contenuti* a partire da video già esistenti nei diversi dataset. Per esempio, l'AI può creare video con stili diversi a partite dal video iniziale, può aggiungere effetti speciali ai video in modo da ottenere nuovi contenuti diversi o può creare nuove scene video basate su istruzioni specifiche inserite in input.
- *La creazione di contenuti di nicchia* che può essere ottenuta attraverso un utilizzo corretto dell'AI Generativa. Quest'ultima non ha limiti di generazione e le possibilità sono molteplici. Infatti, con l'AI un qualunque elemento può essere inserito all'interno di un video che può contenere oggetti che eventualmente non coesistono tra di loro ma che con l'AI Generativa è possibile mettere insieme.

- *Il risparmio di denaro e di tempo* nella produzione dei video in cui si ha la possibilità di non lavorare con un'agenzia o azienda di intrattenimento specializzata nel video editing e nella produzione di video ma bensì ci si affida, per gran parte del lavoro, ai generatori o tool di video AI che comportano solo un costo dal punto di vista del piano tariffario a pagamento che prevedono per effettuare le generazioni dei contenuti. I piani tariffari possono essere mensili o annuali e ogni sottoscrizione pagata offre all'utente una serie di funzionalità. Quindi, la generazione dei video con l'AI porta a un risparmio di denaro, oltre che di tempo, in quanto ci si affida a dei tool e non a membri di aziende esterne che lavorano nell'ambito di produzione video e che necessitano di un pagamento ben più rilevante rispetto alla cifra che si spende per i tool AI di Video Generation.
- *L'integrazione dei video* fatti in AI con i software di video editing esistenti attraverso la quale gli utenti possono continuare l'eventuale modifica del video generato all'interno di programmi adatti al video editing come Adobe Premiere, DaVinci Resolve e altri software simili. In questo modo, si porta a termine il workflow della produzione del video AI, senza che questo venga alterato dal programma che si decide di utilizzare per ottimizzarlo ulteriormente.
- Il *supporto multilingua* disponibile nei prompt testuali che si possono scrivere in input negli applicativi di AI Generativa. In questo modo l'utente può decidere di scrivere il prompt testuale di ingresso con una lingua anche diversa dall'inglese, perché i tool AI di generazione video sono istruiti anche per comprendere lingue differenti dall'inglese. Nonostante ciò, si predilige l'utilizzo della lingua inglese per la scrittura dei prompt di testo.
- La *collaborazione al lavoro* dove gli utenti, una volta che generano dei contenuti video o delle immagini, possono condividere i loro risultati con la community che utilizza uno specifico tool e possono ricevere feedback per ottimizzare la generazione. Inoltre, gli utenti possono prendere i risultati generati da altri e remixarli. Questo significa che possono modificarli a loro piacimento per generare qualcosa di nuovo in un approccio diverso.

Oltre ai vantaggi, si possono individuare anche degli svantaggi che l'AI generativa di video può presentare. Tra questi emergono:

- La *mancanza di creatività e di originalità* con l'AI che, a volte, è in grado di generare dei video che sono troppo simili a contenuti già esistenti e quindi l'inventiva del nuovo contenuto non è sempre affidabile. Difatti, l'AI ha un range limitato di idee e ispirazioni e si ha l'assenza dell'intuizione umana e del pensiero creativo.
- Il *pericolo di diffondere disinformazione* con la possibilità di uso improprio dei video generati per manipolare informazioni e, successivamente, diffondere notizie false. Per esempio, l'AI può essere usata per creare video realistici deepfake attraverso la modifica del volto degli attori che appaiono nelle diverse scene e in cui si nota il rischio, tramite questa generazione, che i video possano seriamente danneggiare la reputazione di specifici individui che vengono mostrati al posto dei soggetti originali del video ma questo vale anche nel senso opposto nei confronti dell'attore originale che è stato sostituito da una nuova figura.
- La *perdita di posti di lavoro* per i video maker e animation artists. Quindi, l'AI Generativa può sostituire il ruolo lavorativo che queste figure hanno e questo può portare le aziende ad assumere meno dipendenti specializzati nel settore multimediale del montaggio e della produzione di video, siano essi di tipologia live-action, animazioni 3d, 2d o altro.
- I *problemi legali e di copyright* eventualmente relativi ai contenuti generati che possono portare alla violazione delle leggi del diritto di copia e ai diritti di proprietà intellettuale, soprattutto se i video generati dall'AI partono da video che appartengono a un certo soggetto o ente. Questa problematica è però poco diffusa e rara da trovare.

4. Ricerca e analisi condotta sui tool AI generativi

4.1 Tool AI non commercializzati

Nella vasta scelta di strumenti o tool di AI Generativa, rientrano quelli che sono stati sviluppati e implementati ma non ancora rilasciati ufficialmente nel mercato digitale. Gli applicativi di Generative AI non ancora commercializzati presentano una loro struttura interna e offrono, a chi ha accesso al loro utilizzo, un certo numero di funzionalità che sono comuni a quelle di altri tool AI generativi che sono stati già ampiamente rilasciati da diverse aziende produttrici.

Nelle implementazioni recenti, rientrano tra gli strumenti AI generativi non commercializzati quelli di proprietà di aziende come Meta, Google, Nvidia e altre simili. Infatti, i tool di queste aziende sono stati principalmente testati ed utilizzati in maniera pratica dai dipendenti delle stesse. Dunque, le figure come ingegneri, ricercatori, tecnici e personale specializzato nell'AI Generativa delle aziende hanno avuto l'occasione di provare più volte questi strumenti sviluppati ma non ancora commercializzati agli utenti online.

La scelta delle aziende, di non pubblicare in maniera aperta gli applicativi di AI Generativa, può essere di due tipi:

- *Momentanea*: in quanto si attendono di definire gli aspetti burocratici e di privacy per far valere i diritti dell'azienda creatrice dell'applicativo ma una volta risolti questi ultimi aspetti, i tool vengono rilasciati in commercio senza particolari problemi.
- *Definitiva*: questa risiede nel fatto che l'azienda produttrice del tool decide di utilizzare lo strumento per scopi di ricerca interna senza dare la possibilità agli utenti esterni di fruire delle funzionalità e delle possibilità generative che quel determinato applicativo consente di fare.

4.1.1 Il caso Meta: logica metodo EMU Video

Tra le aziende che hanno realizzato dei tool AI Video Generativi non commercializzati, emerge in maniera tangibile Meta.

Il metodo o modello AI creato e adottato dal team GenAI di Meta, per la generazione di video AI, presenta il nome di EMU Video.

EMU Video è un metodo applicato da Meta per l'AI Video Generation che si basa su un modello fattorizzato e matematico che segue uno schema ben preciso e strutturato.

Nel dettaglio, il metodo rappresenta una ricerca e una scoperta condotta sull'AI Generativa e ragiona su un approccio di fattorizzazione della generazione AI di tipo Text-to-Video. Questo approccio è possibile rispettarlo tramite un condizionamento esplicito dell'immagine (o frame video) che si ottiene ad ogni step della generazione video che viene prodotta.

Per fattorizzazione si intende quel processo di scomposizione del video che si desidera generare, in tante immagini inserite in successione.

EMU Video si basa su dei modelli di diffusione presenti nei dataset dell'azienda proprietaria e fattorizza o segmenta la generazione video in 2 step:

- Prima agisce sulla generazione dell'immagine (o frame video) condizionata ed ottenuta dal prompt testuale inserito in input;
- Successivamente si passa alla generazione del video in sé, condizionato ed ottenuto dal prompt testuale dato in ingresso e dalle immagini generate a partire dal testo. Queste vengono inserite in serie progressivamente per ottenere il video in output e, dunque, l'animazione.

La fattorizzazione della Text-to-Video AI Generation guidata dal condizionamento esplicito dell'immagine creata, si appoggia su un modello logico-matematico-operazionale che è illustrato di seguito:

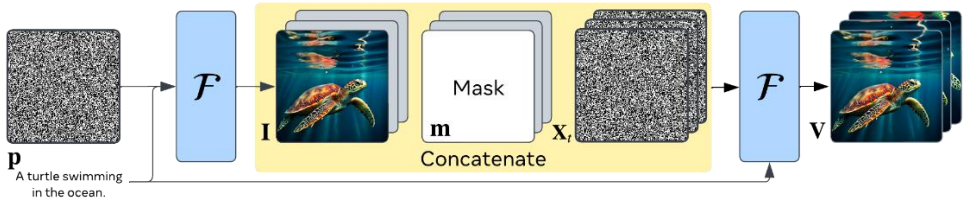

Figura 4.1 Schema operazionale con applicazione del metodo EMU Video per ottenere la generazione del video

Il grafico del modello EMU, o della generazione AI text-to-video fattorizzata, implica prima una generazione di un'immagine I condizionata che dipende dal testo p inserito in input. Successivamente, si applica un condizionamento più forte, basato sull'ultima immagine generata e sul testo inserito in ingresso, per ottenere gli altri frame video e, dunque, le altre immagini successive che costituiscono il video V da generare. Il video V sarà formato da una serie di immagini messe in successione.

Per applicare il condizionamento con il modello EMU (denominato F) sull'immagine generata, si porta a 0 l'immagine temporaneamente (quindi questa diventa bianca con i bit a 0) e si effettua la concatenazione tra l'immagine bianca e la maschera binaria (composta da bit 0 e 1). La concatenazione consente di individuare quali frame sono a 0 e quali presentano il rumore di input che è stato inserito all'inizio del processo di generazione, cioè quando si è scritto il testo di input p.

Dall'approccio procedurale della figura inserita sopra, si evince che l'obiettivo della generazione text-to-video è quello di costruire un modello di diffusione latente F (l'elemento cardine di EMU) che ha in input un prompt testuale p che tende a generare un video V, composto da frame video RGB (denominati T) che appaiono in output. Durante l'applicazione del modello F, si effettua al suo interno l'operazione della concatenazione citata precedentemente che serve ad elaborare l'immagine o il frame video con il rumore.

In particolare, spiegando il metodo o modello EMU Video in modo più tecnico e approfondito, si denota che il modello F (che poi sarebbe l'EMU) viene inizializzato con un modello già pre-istruito di generazione text-to-image e questo serve a garantire che il modello sia capace di generare immagini al momento dell'inizializzazione quando si legge il testo p di input e, quindi, di ottenere il primo frame I. Pertanto, bisogna istruire solamente il modello F per risolvere il secondo step, ovvero estrapolare un video condizionato da un prompt testuale e dall'immagine (o frame) iniziale generata.

Si addestra il modello F usando coppie video-testo tramite il campionamento del frame iniziale I e attraverso la richiesta, che si esegue al modello F, di predire i frame T del video usando il condizionamento di entrambi gli elementi: il prompt p e l'immagine iniziale generata I.

La generazione dei frame T messi in successione (le diverse immagini inserite una dopo l'altra) portano ad avere il video V formato da frame T RGB con certe dimensioni spaziali. Dunque, il video di output adotta spazi e modelli di colore RGB.

Dal momento in cui si adottano modelli di diffusione latenti (quindi che implicano una certa propagazione nel tempo), all'inizio del processo, si converte il video V in uno spazio latente X usando un autocodificatore dell'immagine applicato sulla larghezza del frame che riduce le dimensioni spaziali del frame per facilitarne la generazione.

Lo spazio latente può essere riconvertito nello spazio dei pixel iniziali usando il decodificatore dell'autocodificatore. I frame T del video, con questo step di ritorno, sono rumorizzati/disturbati indipendentemente per produrre l'input rumorizzato Xt (segnale di rumore/disturbo) a cui il modello

di diffusione latente è istruito per derumorizzare (rimuovere il rumore) e, quindi, togliere il rumore per avere il video V senza noise e con frame T elaborati in maniera corretta.

Nel processo è fondamentale che venga trattato e inserito del rumore per consentire al modello una generazione fattorizzata migliore, anche perché avere dei frame T intatti, senza che prima posseggono del rumore, può essere di difficile elaborazione e, soprattutto, attraverso il rumore si riduce la dimensione del frame di output T, che è stato elaborato prima di essere generato a video.

Questo processo viene ripetuto fino a quando non si ha un numero di immagini sufficientemente valido per garantire una buona animazione e quindi un buon video di output.

Rispetto ad EMU Video, altri metodi AI recenti generano i frame video T contemporaneamente usando il condizionamento basato sul solo testo inserito in input e applicando la pura generazione T2V, con un grande uso di modelli di diffusione.

La semplicità nell'implementazione del modello EMU porta a considerare il fatto che il metodo può essere istruito usando dataset standard di text-video, i cui contenuti prendono spunto da questi e il metodo non richiede una vasta cascata di modelli di diffusione da utilizzare all'interno della generazione AI del video. Infatti, un numero valido di modelli può essere pari a 7, che sono sufficienti per generare video ad alta risoluzione, con una qualità decente e una fedeltà al prompt testuale molto buona.

EMU Video è un metodo piuttosto semplice perché semplifica il lavoro di generazione video in quanto, tramite i due step sopra citati relativi al passaggio da testo a immagine e poi da immagini + testo a video, si notano risultati migliori perché progressivamente si ottengono delle immagini intermedie che aiutano a generare il video in maniera ottimale. Ad ogni step, si generano immagini progressive che, unite insieme attraverso una successione, compongono il video. Le immagini fungono da singoli frame video e, ad ogni step, si tiene conto del testo inserito in input e delle immagini ottenute nei passaggi precedenti.

Nel metodo EMU Video si generano tante immagini intermedie fattorizzate e la fattorizzazione gestisce il movimento tra le diverse immagini ottenute e, dopo una successione di 4/5 frame video, si ottiene l'effetto di animazione completa tramite la serie rapida di immagini create che conduce alla creazione del video finale.

Inoltre, EMU Video si basa sull'ipotesi che il condizionamento più forte per la generazione del video AI si affida al testo e all'immagine generata e questa doppia condizione (o dipendenza) può migliorare la generazione del video che sarà rappresentata da un certo numero di immagini posizionate in successione e che saranno elaborate per dare risultati migliori nel video di output. Quest'ultimo conterrà una buona qualità (o risoluzione) e presenterà un'elevata fedeltà al prompt testuale inserito in input.

In altri modelli AI, invece, avviene una scrittura del testo in input che conduce alla generazione dell'immagine e su di essa si esegue una predizione dei movimenti per ottenere il video finale. Inoltre, si manifesta un utilizzo di tanti modelli di diffusione sull'immagine iniziale ottenuta. Rispetto agli altri metodi di AI Video Generation, che richiedono una grossa quantità di modelli di diffusione, EMU Video richiede l'utilizzo di pochi diffusion models per generare, allo stato attuale del metodo, video a 512 pixel (px) che hanno una durata di 4 secondi e un frame rate pari a 16fps.

Dal punto di vista pratico, il team di Meta ha confrontato i risultati ottenuti da una generazione video eseguita con il metodo EMU Video, con una generazione video prodotta dai metodi AI utilizzati dai tool commercializzati e accessibili agli utenti.
Il confronto è stato condotto basandosi su due parametri:

- Qualità complessiva del video generato;
- Fedeltà al prompt testuale inserito in input.

Per il confronto tra le due tipologie di modelli, è stato utile attingere da una serie di feedback e suggerimenti ricevuti da valutatori umani che hanno dovuto scegliere i video AI più convincenti, basandosi sui due parametri citati precedentemente.

Dal confronto, si evince come il metodo EMU Video sia stato migliore rispetto ai modelli dei tool commercializzati (tra cui sono presenti tool come Runway con il metodo Gen-2 e PikaLabs con il metodo Pika). Quindi, i video a 512 pixel con durata di 4 secondi e frame rate di 16fps hanno avuto un impatto migliore, nei 2 parametri considerati, rispetto ai video generati dai modelli AI relativi ai tool rilasciati in commercio (commercializzati).

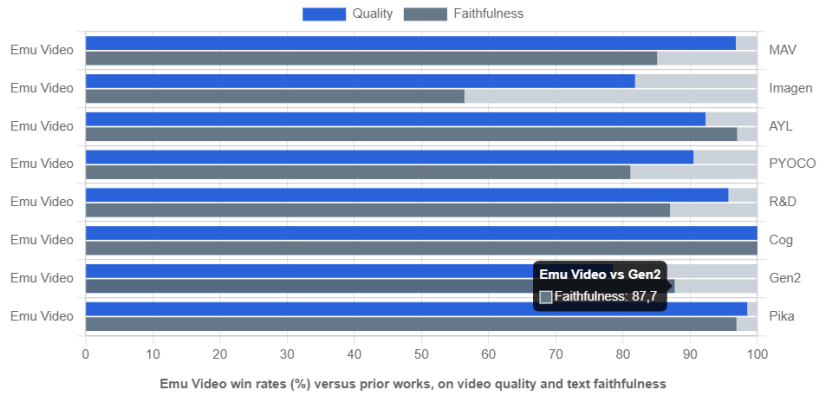

Figura 4.2 Grafico di confronto, basato sui 2 parametri, tra il metodo EMU Video e i modelli (metodi) dei tool AI Video Generativi commercializzati (in basso si nota Runway con metodo Gen2 e PikaLabs con metodo Pika)

Relativamente ai confronti e ai risultati emersi è utile considerare che è stata adottate una valutazione umana robusta (metodo di valutazione JUICE). Il modello EMU VIDEO è stato valutato da valutatori umani tramite dei giudizi basati sui due parametri di qualità e fedeltà dei video generati in output.

Il termine della tecnica di valutazione JUICE proviene dal fatto che il team di Meta ha rivolto la richiesta di voto a dei valutatori umani che hanno giustificato (JUstify) le loro scelte (choICE) quando hanno scelto una generazione video rispetto a un'altra. Questo approccio ha offerto ai diversi sviluppatori una base valida di valutazione.

I giudizi sono stati assegnati sulla qualità del video intesa in termini di: nitidezza dei pixel dei frame video; fluidità dei movimenti; oggetti e scene riconoscibili; coerenza dei frame e quantità di movimento (GENERAL MOTION). Mentre per fedeltà al prompt testuale si è inteso l'allineamento del testo spaziale e l'allineamento del testo temporale.

Per concludere, oltre alla tipologia di Text-to-Video Generation, il metodo EMU Video può essere applicato per eseguire Image-to-Video Generation e Text+Image-to-Video Generation.

4.2 Tool AI commercializzati

Nel precedente paragrafo sono stati trattati gli strumenti di AI Generativa non commercializzati, con un focus specifico sul contesto video tramite il metodo EMU. Invece, questa sezione si focalizza sugli applicativi AI generativi rilasciati in commercio e, tra questi, viene considerato il tool Runway. Questo è stato scelto come principale esempio di strumento di Generative AI che è stato sviluppato, rilasciato in rete e commercializzato per essere fruibile dagli utenti e che garantisce differenti tipologie di generazioni.

4.2.1 Il caso Runway: logica di funzionamento

Tra i tool di AI Video Generation rilasciati in commercio, Runway rappresenta un esempio concreto di come l'AI Generativa si stia evolvendo in ambito video e immagini. L'evoluzione della produzione e della generazione video guidata dall'AI induce a considerare questo applicativo come un esempio forte di innovazione e funzionalità video generative avanzate.

Runway è stato scelto come caso studio per capire la logica di funzionamento presente al suo interno. Inoltre, una considerevole porzione di questa logica è comune anche ad altri applicativi video generativi.

La logica di Runway, come quella di altri tool AI video commercializzati simili (Pika, Genmo, Kaiber e altri), si basa su numerosi modelli di diffusione latente che sono classificabili come modelli di ML (Machine Learning) e questi si appoggiano su dei framework ML esistenti. Infatti, Runway presenta nella sua denominazione originale la sigla ML e da questa osservazione si estende il nome completo RunwayML che testimonia come il tool si appoggi su modelli e algoritmi di ML. Nella logica back-end di Runway, emerge la serie di framework ML implementati: TensorFlow, PyTorch e Keras.

L'operazione di sintesi o generazione AI dei video si basa sui framework ML citati ma anche su altri. Una volta compreso il funzionamento dei 3 framework di back-end, si comprende come i tool AI video commercializzati agiscono in generale, compreso l'applicativo Runway.

I contenuti che vengono generati all'interno di Runway, ma anche in altri tool AI di generazione video messi in commercio, vengono definiti come modelli ML personalizzati, in quanto questi contenuti, per essere elaborati, si affidano ai framework ML implementati.

La logica Machine Learning di Runway implica a definire i contenuti come se fossero dei modelli ML personalizzati, siano essi generati da zero mediante le funzionalità offerte dall'applicativo oppure generati a partire da modelli o contenuti esistenti, prodotti da altri utenti e su cui è possibile effettuare operazioni di remix per ottenere altri contenuti/modelli ML.

Una delle caratteristiche principali dei tool AI video generativi rilasciati, e di Runway versione Gen-2, è che si affidano a dati esistenti per generare nuovi contenuti. I dati esistenti vengono estrapolati dagli enormi dataset a cui l'azienda di sviluppo del tool ha accesso. I dataset di contenuti relativi ai tool AI video commercializzati hanno dimensioni più rilevanti rispetto ai dataset usati dall'applicativo basato sul metodo EMU VIDEO. In contrapposizione, i tool AI commercializzati di video generation utilizzano molti più modelli di diffusione latente, perché la generazione AI dei video agisce sulla stessa singola immagine generata e, su di essa, si effettuano tutte le operazioni di predizione del movimento per generare il video. Invece, con il modello EMU VIDEO si agisce gradualmente su successioni di immagini per generare il video AI di output.

4.2.2 Framework ML di back-end

Negli applicativi di Generative AI rilasciati in commercio, AI e Machine Learning coesistono per generare contenuti tramite l'utilizzo di modelli di diffusione ML specifici, gestiti da framework di back-end. Runway, in qualità di applicativo di AI Video Generation, adotta i framework ML per gestire ed effettuare le generazioni dei video. Oltre ad appartenere al contesto Machine Learning ed AI, questi framework si affidano al Deep Learning, per una certa porzione logica di funzionamento.

I framework ML di back-end che Runway utilizza sono 3: Tensorflow, Keras, Pytorch.

Le definizioni di questi 3 framework dicono che:

- *Tensorflow* è una libreria software di basso livello creata da Google per implementare modelli di ML e per risolvere problemi numerici complessi. Il framework esegue calcoli tramite la conversione di

ogni elemento in una forma grafica. Le variabili del grafico che viene rispettato si chiamano Tensors (tensori) e le operazioni matematiche vengono chiamate operatori.

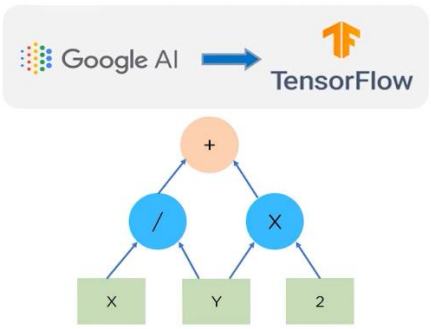

Figura 4.3 Grafico di operazioni matematiche tra tensori e operatori, relativi a Tensorflow (implementato da Google AI)

- *Keras* è un API (Application Programming Interface) di deep learning ad alto livello, scritta in Python per garantire una facile implementazione e computazione di reti neurali. Keras si integra con diversi motori di back-end per aiutare a mantenere una facilità di implementazione tramite una veloce computazione delle operazioni.

Figura 4.4 Librerie e framework ML e DL a cui Keras fa affidamento

- *Pytorch* è un API di basso livello sviluppata da Facebook per processare ed elaborare il linguaggio naturale e la computer vision.

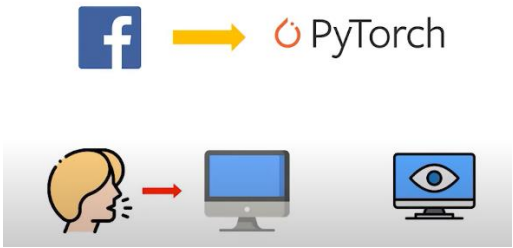

Figura 4.5 Elementi chiave che rappresentano Pytorch (implementato da Facebook)

I 3 modelli o framework ML presentano delle differenze in termini di: livello di API; velocità; architettura; dataset e debugging; facilità di sviluppo; facilità di distribuzione.

Qui di seguito sono mostrate 6 tabelle che sintetizzano i 3 framework ML con i relativi parametri di confronto:

Tabella 4.1 Livello di API

Modello ML	**Livello di API**
Tensorflow	Alto e basso
Keras	Alto
Pytorch	Basso

Tabella 4.2 Velocità

Modello ML	**Velocità delle prestazioni**
Tensorflow	Molto veloce (alte prestazioni)
Keras	Più lento rispetto a Tensorflow
Pytorch	Stessa velocità di Tensorflow

Tabella 4.3 Architettura

Modello ML	**Architettura**
Tensorflow	Complessa e difficile da usare
Keras	Più semplice
Pytorch	Complessa

Tabella 4.4 Dataset e debugging

Modello ML	**Dataset e debugging**
Tensorflow	Usato per modelli di performance molto alte, debugging difficile
Keras	Usato per dataset più piccoli, debugging facile e meno frequente
Pytorch	Usato per dataset grandi, debugging più facile rispetto a Tensorflow

Tabella 4.5 Facilità di sviluppo

Modello ML	**Facilità di sviluppo**
Tensorflow	Difficile da sviluppare e per scrivere codice
Keras	Facile per sviluppare, meglio per neofiti
Pytorch	Più facile da capire rispetto a Tensorflow

Tabella 4.6 Facilità di distribuzione

Modello ML	**Facilità di distribuzione**
Tensorflow	Facile da distribuire
Keras	Intermedia distribuzione
Pytorch	Facile da distribuire ma non quanto Tensorflow

Dalle tabelle parametrizzate dei 3 framework ML, che sono alla base dei tool AI video commercializzati come Runway, è utile sottolineare come l'unione dei 3 modelli implica una collaborazione per garantire generazioni AI video complete. I 3 framework ML coesistono tra di loro ed ognuno di questi è utilizzato, lato back-end, per eseguire procedure ed operazioni specifiche. Sintetizzando:

- *Tensorflow* è implementato per svolgere operazioni logiche e matematiche, fondamentali per la riuscita della generazione AI del video.
- *Keras* serve per facilitare l'implementazione e la computazione a livello di calcolo in modo tale che la generazione video sia realizzata in tempi ragionevoli.
- *Pytorch* è applicato per intervenire sulla comprensione verbale del testo che viene inserito in input nella generazione video e agisce sull'elaborazione grafica del video che si vuole generare, facilitando l'applicazione dello stile scelto al contenuto e accelerando il processo di rendering del video AI.

4.3 I game changer dell'AI Generativa

In questo paragrafo si esprimono quelli che, allo stato attuale dello sviluppo dell'AI Generativa, sono considerati i game changer o pionieri della nuova tecnologia creativa. Questa classificazione di alcuni tool AI generativi, come i principali punti di riferimento per la loro tipologia di generazione di appartenenza, è stata effettuata dal candidato ma è stata anche confermata (tramite ricerca sul web) da altri utenti in rete che hanno avuto la stessa opinione.

Per classificarli, è utile segmentare la Generative AI nelle diverse e principali tipologie di generazione che consente di eseguire: Video Generation; Image Generation; Text Generation; Audio e Voice Generation. Oltre a queste, si stanno sviluppando nuove categorie, con relativi tool AI generativi, che consentono di fare operazioni più specifiche come:

- La rimozione dei watermark dai video generati, per ottenere video senza il marchio del tool AI utilizzato;
- L'aumento della qualità o risoluzione di video e immagini, per ottenere contenuti visivamente migliori (*enhancer tool*);
- La rimozione, ed eventuale sostituzione, di sfondi da video e immagini, per generare contenuti con background diversi e per estrarre l'eventuale sagoma (o silhouette) della figura inserita all'interno del contenuto;
- La rimozione o diminuzione di rumore e disturbi sonori nei segnali audio e nelle registrazioni vocali (*voice and audio cleaner tool*).
  Oltre a queste, sono state diffuse tante altre tipologie di operazioni che l'AI Generativa consente di eseguire per avere output modificati e migliorati.

  Dalle ricerche eseguite e dai lavori effettuati dal candidato per ottenere la generazione di contenuti AI come testi, immagini, video, voci e audio, sono stati scelti alcuni tool come capisaldi della Generative AI, raccogliendoli per categoria di generazione.

  Dunque, per la categoria *Text Generation* è stato individuato il tool chatGPT (versioni 3,5 e 4) come migliore esempio di generazione AI testuale ottenuta a partire da richieste specifiche inserite dall'utente che crea delle chat conversazionali con l'applicativo.
  Per l'*Image Generation*, sono stati individuati applicativi del calibro di Midjourney, DALL-E (versione 2 e 3), Leonardo. Ognuno di questi consente di generare immagini in output a partire da specifici input, utilizzando tecniche differenti di generazione e applicando stili diversi. Le generazioni di immagini che si possono ottenere con questi strumenti sono: text-to-image generation o image-to-image generation.
  Nel contesto della *Video Generation*, i principali candidati come migliori strumenti sono: Runway (con le versioni Gen-1 e Gen-2), Kaiber, Genmo, Pika. Questi offrono generazioni video a partire da diverse tipologie di input che possono essere: testi (prompt testuale), immagini, unioni di testo e immagine, altri video.
  Infine, per la categoria *Audio e Voice Generation* sono stati scelti, come migliori tool AI, ElevenLabs e Synthesia. Entrambi riescono a generare delle voci precise e reali. Inoltre, l'ultimo applicativo citato è utilizzato anche per generare video con la presenza di avatar e a questi si assegnano le voci generate all'interno del tool.

4.4 Modalità di selezione dei tool AI

I tool AI utilizzati per effettuare le operazioni di analisi e generazioni di contenuti (come la realizzazione dei video AI di confronto), sono stati scoperti, individuati e selezionati attraverso diverse azioni mirate e seguendo quelle che sono state le esigenze che il candidato ha dovuto seguire. Per individuare quali applicativi AI generativi utilizzare, sono state svolte queste attività:

- Ricerche effettuate sul web, tramite motore di ricerca, in cui si è svolta una navigazione approfondita di diversi tool AI generativi cercati per tipologia di generazione;
- Newsletter ricevute in azienda utili alla promozione dell'uso di tool AI generativi di diversa tipologia, siano questi già consolidati nel panorama Generative AI oppure applicativi emergenti ma affidabili;

- Visione di video sulla piattaforma multimediale streaming Youtube, in cui esperti della Generative AI hanno dimostrato il funzionamento di alcuni strumenti, definendoli come migliori per specifiche generazioni di contenuti.

  Questi metodi di ricerca sono stati utili per capire e decidere quali tool AI generativi selezionare e adottare per:
- Svolgere le generazioni di video e altri contenuti nel migliore dei modi, relative al progetto di tesi;
- Effettuare analisi approfondite, basate su dei parametri tecnici e non, che evidenziano le funzionalità e le capacità degli applicativi in modo preciso.

4.5 Categorizzazione dei tool AI scelti

Dopo aver individuato e selezionato i diversi applicativi AI generativi, è stato necessario organizzarli per categorie di generazione di appartenenza. Quindi, sono stati assegnati gli strumenti selezionati, alle tipologie di generazione che consentono di svolgere.

Quindi, la suddivisione dei tool AI scelti è stata eseguita attraverso una categorizzazione precisa in cui gli applicativi AI sono stati assegnati alle rispettive tipologie di generazione. Le categorie di generazione considerate per la suddivisione dei tool sono: AI Video Generation; AI Image Generation; AI Voice e Audio Generation.

La categoria AI Text Generation non è stata trattata come categoria di generazione in cui inserire e analizzare determinati tool AI, in quanto è stato scelto a priori un unico tool affidabile ed appartenente a questa categoria: chatGPT (*versioni 3,5 e 4*).

Dunque, gli strumenti AI generativi selezionati sono stati inseriti all'interno di una (o più) di queste categorie di generazione e, per ogni categoria, sono state prodotte delle tabelle che indicano l'analisi tecnica condotta sui differenti tool AI scelti e inseriti all'interno di una nota tipologia di generazione.

4.6 AI Video Generation: ricerca, utilizzo e analisi dei tool AI

Nel lavoro di tesi condotto, la categoria di generazione maggiormente considerata e trattata, tra quelle citate nel paragrafo precedente, è l'AI Video Generation. Relativamente a questa tipologia di generazione, sono stati ricercati, selezionati e utilizzati differenti applicativi AI Video Generativi. La focalizzazione sulla generazione di video AI ha permesso di ottenere una ben definita conoscenza pratica e tecnica degli strumenti video individuati, disponibili online.

Infatti, le nozioni tecniche apprese dall'utilizzo pratico degli applicativi AI video generativi, hanno permesso al candidato di svolgere delle analisi, seguendo dei parametri tecnici, relative allo studio delle funzionalità tecniche offerte da questi tool di AI Video Generation.

Le analisi degli applicativi AI video generativi consistono nell'elaborazione di una serie di tabelle contenenti informazioni e dettagli utili a spiegare i tool dal punto di vista tecnico e funzionale. Le tabelle contengono colonne con i parametri, tecnici e non, rispettati per effettuare l'analisi dei diversi tool AI di Video Generation.

Nelle tabelle, mostrate di seguito, sono presenti i tool AI video generativi che sono stati ricercati, utilizzati e analizzati per realizzare i video AI di confronto del progetto di tesi.

Tabella 4.7 Prima analisi effettuata sui tool AI video generativi

Categoria tool	Input e output consentiti nella generazione	Nome tool	Generazione browser (web app)	Generazione da server Discord	Ris.	Max upscale/upscaling

Video, audio, image Generation	Gen1: video-to-video; Gen2: text-to-video; image-to-video; text+image-to-video	*RunwayML* (Gen1, Gen2)	Si	No	HD, fullHD, 2K;	upscaling attivabile, risoluzioni massime fullHD/2K ris. base è HD (se NON si attiva upscale)
Video Generation	Video Flipbook: text-to-video; text+image-to-video; text+image+audio-to-video; Video Motion: text-to-video; text+image-to-video; Video Transform: video-to-video	*Kaiber*	Si	No	HD, fullHD	Upscaling graduale fullHD e 4K; oppure scegliere subito 4K dal 720p senza passare dal fullHD
Video Generation (browser e Discord)	Da browser: Text-to-video; image-to-video; text+image-to-video; video-to-video; Da Discord: Text-to-video; image-to-video; text+image-to-video	*Pika*	Si	Si (PikaLabs) ed è una versione beta free	Browser: HD, fullHD; Discord: SD, HD	No upscaling su nessuna piattaforma (No browser e discord)
Video Generation	Text-to-video; image-to-video	*Moonvalley*	No	Si	SD, HD	No upscaling
Video Generation	Text-to-video	*Plaiday*	No	Si	SD, HD, fullHD	No upscaling
Video Generation (browser e discord)	Text-to-video; image-to-video	*Neverends*	Si	Si	browser: HD, fullHD e altro; da discord: SD	No upscaling
Image Generation (browser); Motion Generation (browser) Motion=video che mostra immagine animata	Text-to-image; image-to-video (motion)	*Leonardo*	Si	No	Immagini e motion: HD, fullHD;	No upscaling
Video Generation (browser e discord);	Text-to-video; Text+image-to-video; Text-to-image	*Genmo Replay*	Si	Si	Browser: fullHD, 2K; discord: più basse	No upscaling su browser e discord

Image Generation (browser e discord)						

Tabella 4.8 Seconda analisi effettuata sui tool AI video generativi

Nome tool	**Movimenti camera (camera motion)**	**Durata generazione singolo video + estensione**	**Qualità contenuto 3D**	**Effetti post-produzione ed editing video**	**Qualità contenuti live action**	**Frame rate disponibili**
RunwayML	zoom in-zoom out; orizzontale-verticale; pan; tilt; rool	4s+3 estensioni di 4s; durata finale estendibile: tra 16s e 20s	Presente e molto buono; 4 stili 3D generabili	Si presenti, di buon livello, disponibili operazioni di VFX/SFX	Buona	24fps
Kaiber	Flipbook: zoom-in, zoom-out, rotate, up, down, right, left); NO mov camera per Motion e Transform	Flipbook: 3sec-8min; Motion: 3-16sec; Transform: dipende dalla durata video input con vincolo durata max 4min	Discreta, presente denominazione 3D rendering per Flipbook e Transform; NON presente per Motion	No	Buona	12fps Flipbook; 24fps Motion; fps Transform dipende da video input
Pika	pan, tilt, rotate, zoom da browser; da discord gli stessi ma con uso comando - camera	3sec browser e discord; browser estensione 4sec ogni quanto si vuole; discord NO estensione durata	Molto buona browser; discreta discord	No	Buona browser; bassa discord	Da 8 a 24fps browser e discord; frame rate default 24fps browser e discord
Moonvalley	zoom-in, zoom-out, pan left, pan right, pan up, pan down	3 opzioni: breve=1sec; media=2sec; lunga=3sec; durata massima 3 secondi	Discreta	No	Buona	Da 25 a 50fps
Plaiday	Da scrivere nel prompt testuale	Durata base: 3sec; NO estensione durata	Discreta	No	Buona	12fps
Neverends	Da definire nel prompt testuale (browser e discord)	Browser: 3-12sec (no estensione); discord: 3sec (no estensione)	Buona (browser e discord)	No (browser e discord)	Buona (browser e discord)	Browser: 30fps; discord: 10fps
Leonardo	Zoom-in, zoom-out, orizzontale-verticale, vari	4sec (no estensione)	Molto buona	No	Buona	24fps
Genmo Replay	Browser: zoom-in, zoom-out, no-	Browser e discord: 2-4-6sec; no est.	Buona	No	Discreta	15fps

	zoom; rool, pan, tilt, auto Discord: zoom-in, zoom-out, pan-x (right, left), pan-y (up, down)	due piattaforme				

Tabella 4.9 Terza analisi effettuata sui tool AI video generativi

Nome tool	**Watermark**	**Negative prompt**	**Tutorial**	**General motion**	**Aspect ratio**	**Elenco stili visivi/grafici impostati**
RunwayML	No	No	Si presenti	Valori da 1 (molto lento) a 10 (molto veloce)	16:9, ma ne offre altri	Si, presente
Kaiber	No	No (per i 3 tipi di video)	No	Valori da 0 a 10 (per tutti e 3 i tipi di video)	I principali per Flipbook e Motion; Transform ha quello del video di input	Si, presente
Pika	Si discord (free plan); no browser (paid plan); si browser (free plan)	Si browser; no discord	No browser, si discord	Browser: valori da 0 a 4; discord: definire nel prompt tramite parametro -motion	Browser: principali formati video; discord: definirlo nel prompt (--ar 16:9, etc)	No, devono essere definiti nel prompt prima della generazione (browser e discord)
Moonvalley	Si versione free; no paid plan	Si con parametro negative_prompt	Si presenti	No	Non presente, di base 16:9	Si, presente
Plaiday	Si	No	Si presenti	No	Da definire nel prompt testuale	No, devono essere scritti nel prompt prima della generazione
Neverends	No browser; si discord	No browser; si discord con parametro negative	No browser, si discord	Browser: valori da 1 a 5; no discord	Da browser: principali formati video; da discord: parametro aspect-ratio	No, da scrivere nel prompt testuale (browser e discord)
Leonardo	No	Si	No	Valori da 1 a 10	Lo stesso dell'immagine di input (16:9, 1:1, 9:16, altri)	Si, presente

Genmo Replay	Browser: si versione free; no paid plan; discord: no	No browser; si discord con parametro negative_prompt	No browser, si discord	Browser: valori da 50% a 99%; Discord: default 60%, valori da 20% a 99%	Browser: formato auto, 16:9, 9:16, 2:3, 3:4, 1:1, 4:3, 3:2; discord: gli stessi	Da browser: presente tramite voce FX; da discord: presente tramite parametro style

4.6.1 Approfondimenti di alcuni tool

Per 4 dei tool AI video generativi analizzati nelle tabelle, sono stati trattati degli aspetti relativi al Business Model, che unisce due concetti di ricerca relativi alla presenza del Watermark nei video generati e alla possibilità di avere l'opzione di Free Preview, prima del download del video generato, all'interno dell'applicativo. Inoltre, nell'attuale paragrafo, sono presenti approfondimenti e informazioni aggiuntive che spiegano nel dettaglio alcuni applicativi.

1) *Runway* presenta un suo Business Model con piani tariffari che si dividono in mensili o annuali (quest'ultimo solitamente risulta essere quello più vantaggioso in termini di costo complessivo). Dal punto di vista delle funzionalità e dei servizi offerti dai piani mensili e annuali non si hanno cambiamenti nel contenuto, per un piano con la stessa denominazione. L'unico aspetto che varia è il prezzo del piano (per esempio, i piani Standard mensile e annuale hanno le stesse features offerte ma variano nel costo). Oltre ad avere un piano mensile o annuale con la stessa denominazione, col prezzo che varia e con le caratteristiche tecniche uguali, sono evidenti le differenze delle caratteristiche tra i piani con denominazioni diverse (per esempio, il piano Standard cambia dal piano Pro e quest'ultimo cambia dal piano Unlimited e dal piano Enterprise in termini di caratteristiche tecniche offerte).

   Alcune differenze di features tra i piani sono:

- Le memorie in GB degli assets in cui si salvano i contenuti generati;
- Il numero di crediti a disposizione per effettuare le generazioni dei video;
- I tipi di formati di export video;
- Le funzionalità aggiuntive legate ai VFX.

   Runway offre un piano free che non comporta un pagamento ma ha limitazioni tecniche con features ridotte. Al contrario, i piani a pagamento garantiscono diverse funzionalità tecniche ma necessitano di un costo mensile o annuale. Questo è un aspetto valido e universale per tutti i tool AI video generativi finora prodotti e rilasciati.
   Inoltre, se con RunwayML si sottoscrive un piano tariffario, allora le Preview pre-download dei video sono garantite, mentre nel piano free non sono disponibili.
   Infine, rilevante è anche la relazione tra il piano tariffario e la presenza del Watermark: se si paga un piano, allora il watermark scompare dal video generato e la risoluzione del video tende ad aumentare in quanto il watermark viene rimosso e, con esso, scompare la schermata trasparente su cui il watermark è applicato, che causa al video una minore risoluzione e toglie dettagli visivi. Il pagamento implica un miglioramento nella risoluzione perché la schermata del watermark e lo stesso watermark scompaiono e, dunque, il filtro granuloso della schermata del watermark sparisce e si ottiene un video con una maggiore risoluzione.

2) Il tool *Kaiber* presenta 3 tipologie di piani tariffari diversi: Explorer, Pro e Artist. Questi possono essere mensili e annuali e, inoltre, un piano mensile e annuale con la stessa denominazione sono identici in termini di caratteristiche tecniche offerte, mentre differiscono per il costo.
   Nel passaggio tra i diversi piani, oltre ai prezzi, variano anche le features tecniche offerte.

Alcune caratteristiche tecniche proposte tra un piano e l'altro variano:

- Per la durata di generazione disponibile quando si generano i video (totale di 1 minuto o 8 minuti di durata video);
- In base al piano che si possiede si ha la possibilità di effettuare l'Upscaling, oppure no, fino a specifiche risoluzioni;
- Al passaggio tra un piano e un altro, di diverso nome, variano le capacità di calcolo ed elaborazione del tool AI video generativo.

La gestione Preview varia in base al tipo di video che si genera e il pagamento di un piano tariffario comporta la presenza delle preview dei video AI generati.
Infine, per quanto riguarda il Watermark si ha la stessa logica definita per Runway.
Nell'approfondimento relativo a Kaiber, si possono individuare 3 tipologie di video AI che il tool consente di generare. Queste sono:

- Flipbook: video caratterizzato dall'animazione frame by frame;
- Motion: video che presenta una fluidità nell'animazione;
- Transform: video di output che mostra uno stile diverso rispetto a quello del video di input.

3) L'altro tool AI video generativo approfondito è *Pika*. Questo prevede 4 piani tariffari: uno Free denominato Basic e 3 a pagamento denominati Standard, Unlimited, Pro. I 3 piani a pagamento possono essere mensili o annuali, con prezzi che variano per lo stesso tipo di piano (prezzi piani annuali più convenienti dei mensili) ma con features uguali per entrambi i piani mensili ed annuali.

Nel passaggio tra un piano e l'altro (Standard-Unlimited-Pro mensili o annuali) cambiano le features tecniche come:

- Memoria assets;
- Numero di tentativi di generazione video;
- Minuti generabili;
- Velocità di calcolo e quantità di crediti disponibili;
- Gestione Preview dei video, prima del download, diversa dai tool definiti in precedenza in quanto non sono disponibili su entrambe le piattaforme (da browser e da Discord) sia se si sottoscrive un piano che non;
- Gestione Watermark dipende dal piano che si ha a disposizione (se free o pagamento) e con la logica della risoluzione video, che peggiora o migliora, spiegata per RunwayML.

Approfondendo l'utilizzo di Pika, il candidato ha testato delle funzionalità del tool su Discord. Lo strumento funziona con la logica simile al tool AI Midjourney. Infatti, nella fruizione di Pika da server Discord, l'utente deve scrivere una serie di comandi, all'interno della barra di generazione, preceduti da "/" e, dopo questo elemento, si inseriscono dei comandi, con relativi parametri aggiuntivi, per indicare un differente tipo di operazione che si desidera svolgere e che è relativa alla generazione del video che si vuole ottenere.
Tra i comandi sono rilevanti quelli relativi al tipo di formato video che si vuole generare e il tipo di movimento di camera o di motion camera che si vuole ottenere. Oltre a questi sono presenti altri comandi che è possibile svolgere. Il comando principale è /create, seguito dalla scrittura del prompt testuale che indica il contenuto da generare nel video e, successivamente, si possono inserire, all'interno della stessa riga di scrittura della generazione, gli altri comandi che definiscono i dettagli più tecnici del video che si vuole produrre. Quindi, per una generazione video è importante partire dal comando "create" e inserire, in seguito, altri comandi aggiuntivi per avere una migliore riuscita del video (per esempio, il comando relativo al motion camera). A specifici comandi scelti corrispondono dei parametri o delle opzioni per definire meglio cosa si

vuole ottenere. Dunque, scelto il comando del motion camera (-camera), allora sarà possibile scegliere e cliccare, tra le opzioni offerte dal comando, il tipo di movimento di camera che si vuole adottare (zoom-in, zoom-out, pan e altri).
Il tool, tramite il server Discord, consente l'utilizzo di bot (e chatbot) per generare contenuti video AI sia all'interno dei diversi canali pubblici che all'interno del canale privato che l'utente apre per interagire direttamente col bot, senza essere interrotto dalle generazioni di altri utenti che sono presenti nei canali pubblici del server. In questo modo, nel canale privato, si instaura una comunicazione diretta e riservata tra l'utente e il bot/chatbot, che fornisce la possibilità di effettuare la generazione del video AI.
Dunque, se si generano dei video AI all'interno dei canali pubblici, allora si ottengono generazioni pubbliche, in quanto i diversi contenuti generati, con l'ausilio del bot AI, sono visibili anche agli altri utenti. Le generazioni video, generate nei canali pubblici del server Discord di Pika, vengono unite a quelle degli altri utenti e, se l'utente vuole individuare una specifica generazione effettuata in passato, deve svolgere un lavoro di ricerca minuzioso, tramite i filtraggi inseriti nella barra di ricerca presente nel canale.
Invece, se le generazioni dei video AI vengono effettuate nel canale privato (tramite chat privata con il bot AI), allora si ottengono generazioni private e la ricerca di una specifica generazione video all'interno del canale privato è più veloce ed è semplificata dal fatto che non sono presenti le altre generazioni prodotte per gli altri utenti, in quanto si tratta di una chat privata tra l'utente e il bot/chatbot AI di Pika e le generazioni video AI presenti sono solamente quelle richieste dall'utente al bot.

Alcuni comandi importanti per generare i video AI sono:

- - ar oppure -- ar che definisce il formato del video;
- -motion che definisce la velocità del movimento di camera;
- -camera che definisce il movimento di camera;
- Altri parametri spiegati nella documentazione del server discord.

Gli stessi comandi si trovano anche sulla versione da browser del tool Pika con la differenza che i comandi, e i relativi parametri, possono essere selezionati mediante i bottoni dell'interfaccia utente proposta dall'applicativo. Dunque, non c'è la necessità da parte dell'utente, di scrivere i comandi in maniera esplicita ed espressiva nel prompt testuale. Questo significa che, invece di scrivere il comando "/create", è opportuno scrivere il prompt testuale all'interno del campo di scrittura e cliccare sul button relativo alla generazione e creazione del video che è l'equivalente del comando create. Oppure, invece di scrivere "motion camera such as zoom-in" nel prompt testuale, si seleziona il bottone del logo della camera per selezionare il movimento di camera che si desidera inserire nella generazione AI del video.

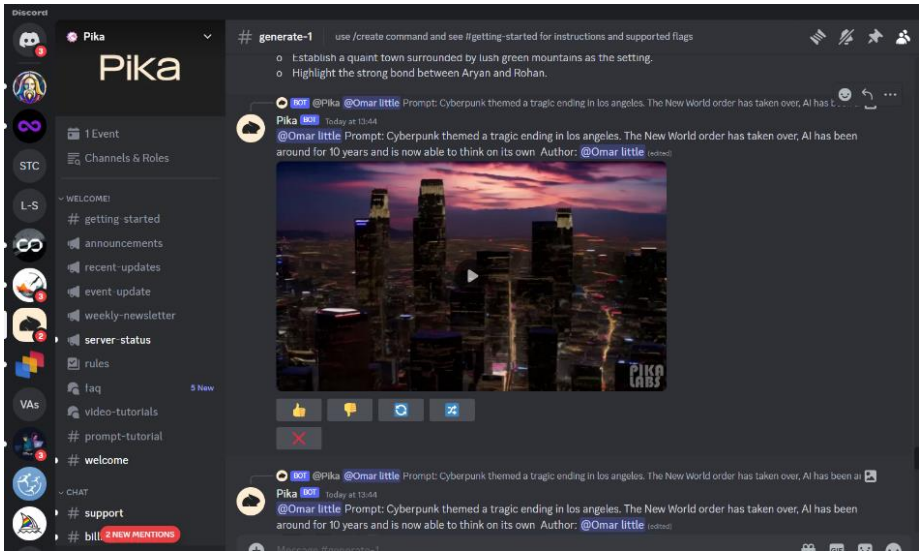

Figura 4.6 Server Discord di Pika con il canale pubblico "generate-1" aperto e contenente generazioni video AI di diversi utenti

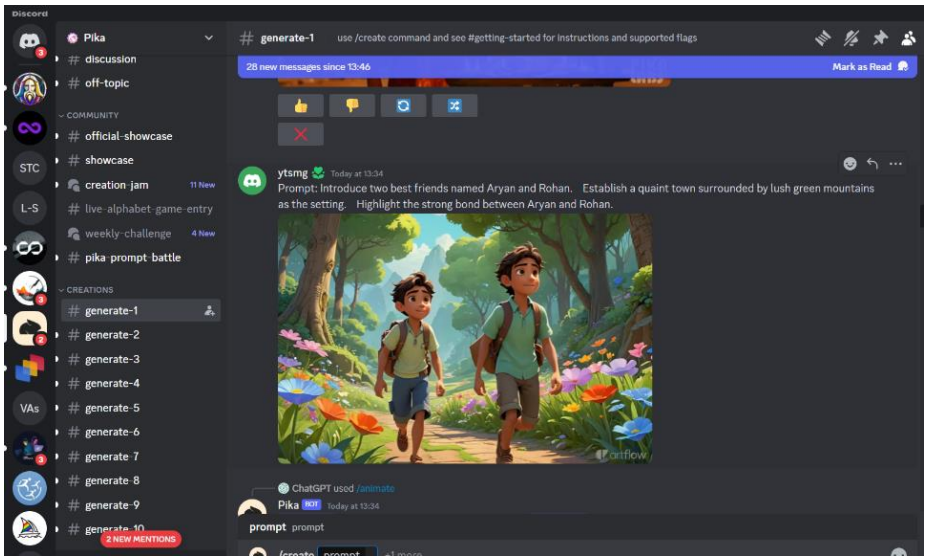

Figura 4.7 Server Discord di Pika e click sul canale "generate-1" con scrittura del comando "/create" per iniziare una nuova generazione AI del video a partire da prompt testuale

4) L'ultimo tool approfondito è Genmo Replay. Genmo offre un piano free e un piano a pagamento sulla versione da browser. Il piano a pagamento è mensile e, rispetto al piano free, contiene più features tecniche.
La presenza del Watermark nei video AI generati è prevista con la versione gratuita del tool, mentre nella versione sottoscritta dell'applicativo non compare nei video prodotti.
Da browser, le Preview dei video, prima del loro download, non sono disponibili in entrambi i piani (free e a pagamento).
Da server Discord, il tool non offre le Preview dei video AI generati. Il Watermark, nei video AI prodotti, non è presente ma si ottiene, come punto a sfavore, una risoluzione dei video generati più bassa. La fruizione del tool da Discord è gratuita.

L'approfondimento condotto, porta a considerare il tool Genmo come uno strumento abile a generare video e immagini con uno stile ben orientato soprattutto verso il 3D.
Il tool, rispetto ad altri, consente di gestire, da browser, il numero di creazioni di video AI che si vogliono ottenere per ogni generazione eseguita. Infatti, il tool presenta gli elementi 1x, 2x, 3x che indicano rispettivamente il numero di video che si generano a seguito della pressione del

tasto “Submit”. Quindi 1x indica 1 video generabile, 2x significa 2 video generabili e 3x rappresenta 3 video generabili ad ogni generazione.
Su entrambe le piattaforme (browser e discord), l’applicativo consente di mettere in loop il video generato e, dunque, l'utente può scegliere se abilitare la riproduzione del video in loop o no, tramite il bottone sulla UI del browser oppure tramite il parametro “loop” che si presenta durante la scrittura del prompt testuale sul server discord.

Quando si genera un video da Discord, tramite il comando “/video”, dopo il prompt testuale si può decidere di inserire il parametro image che consente di dare in input un'immagine e di unirla al testo scritto per generare in output il video AI. Dunque, il video lo genero in output attraverso il comando “/video”. Dopo questo comando, l'utente ha libertà di scegliere se ottenere il video in uscita attraverso un prompt testuale dato in input con aggiunta di diversi parametri oppure può produrre il video inserendo in ingresso l’unione di prompt testuale e immagine più parametri aggiuntivi.

Infine, Genmo consente di generare anche immagini a partire dal testo dato in input. L'immagine la genero tramite il comando “/image” seguito dalla scrittura del prompt testuale inserito in ingresso.

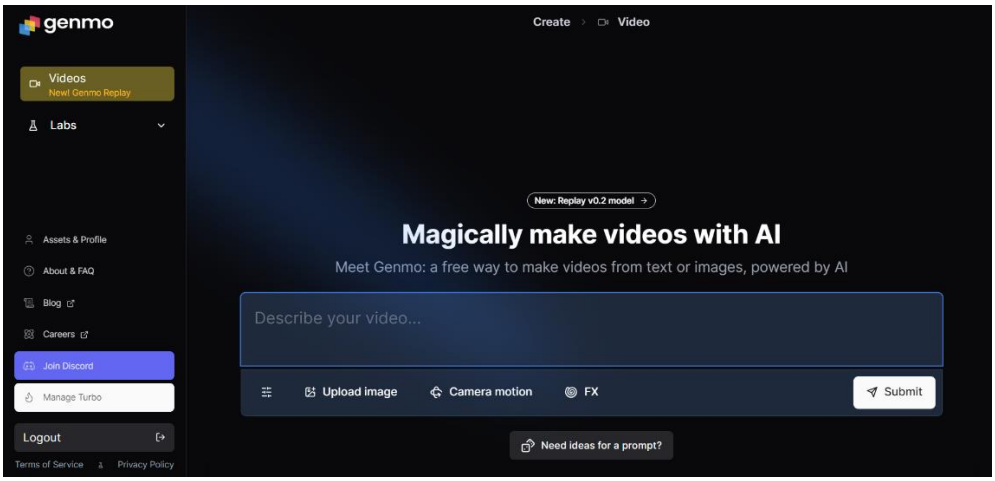

Figura 4.8 Homepage del tool Genmo da browser

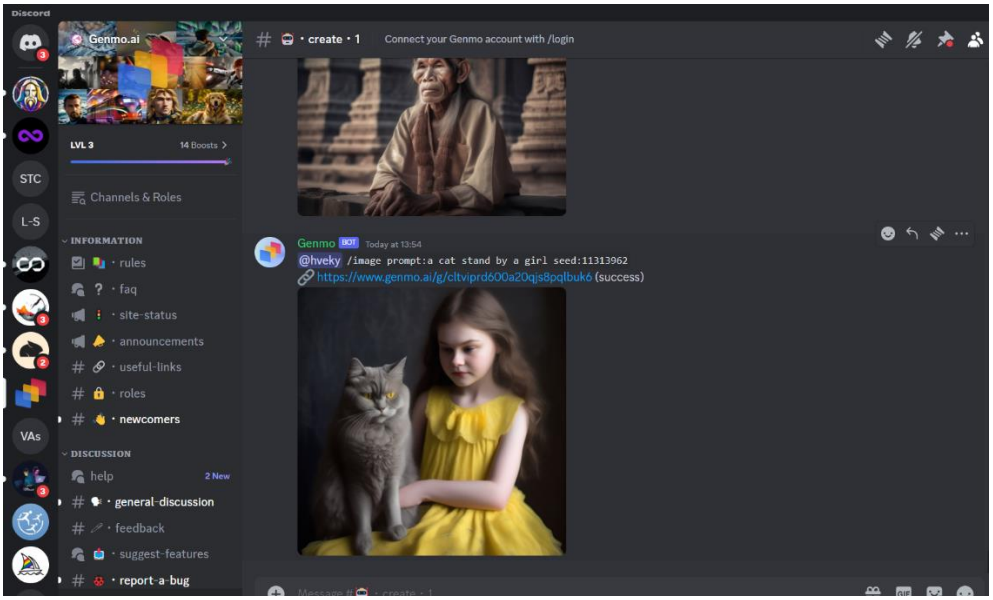

Figura 4.9 Server Discord relativo al tool Genmo con il canale pubblico "create" aperto per effettuare generazioni AI dei video

4.7 Text-to-Video Generation: test e confronti dei tool AI

Una parte del progetto di tesi condotto dal candidato è identificata da una serie di video AI generati da 8 diversi tool AI di Video Generation che, attraverso la tipologia text-to-video generation da loro offerta, hanno mostrato e prodotto risultati differenti in output. I video AI ottenuti sono stati valutati sia soggettivamente dal candidato che oggettivamente da un numero di utenti, tramite le risposte date a un form. Da queste valutazioni espresse, si evince quali video AI sono stati i migliori e i preferiti in termini realizzativi e, inoltre, con i video AI generati, valutati e confrontati tra di loro, si è capito quali sono stati gli applicativi, tra gli 8 considerati, migliori per le generazioni text-to-video effettuate.

Dalle due tipologie di valutazione eseguite (soggettiva e oggettiva) sui video AI ottenuti, sono state espresse due classifiche:

- Una soggettiva, con le considerazioni e le valutazioni date dal candidato, che ha espresso una precisa posizione, nella classifica, per ogni tool AI video generativo utilizzato;
- Una oggettiva, i cui tool AI video generativi sono stati classificati in base ai risultati e alle valutazioni ottenute dal panel, compilato da un certo numero di utenti.

In entrambe le classifiche, in base al piazzamento ottenuto da un certo applicativo AI video generativo, si individuano i tool AI di generazione video più o meno migliori e, di conseguenza, si valutano i video AI generati, relativi ai tool posizionati in classifica, che sono stati più o meno apprezzati.

Per esempio, se RunwayML si posiziona al primo posto, allora significa che è stato il tool maggiormente apprezzato soggettivamente o oggettivamente e, di conseguenza, i video AI generati col medesimo applicativo sono stati definiti come i migliori complessivamente.

Gli 8 applicativi AI video generativi, utilizzati in questa fase pratica, sono quelli inseriti ed analizzati all'interno delle tabelle spiegate e illustrate nel paragrafo 4.6.

Questa fase di generazione dei video AI ottenuti dagli strumenti generativi, dimostra come la text-to-video generation è una valida tipologia di generazione di video AI, purché si utilizzino i giusti elementi in input, fondamentali per la corretta e ottimale creazione AI dei video.

Gli elementi di ingresso o di input a cui si fa riferimento sono i prompt testuali che sono stati scritti e inseriti, seguendo un certo criterio, per ottenere generazioni di video AI in output. In sintesi, più il prompt testuale è scritto e definito meglio, maggiore è la possibilità di ottenere un video, in uscita, di qualità, ben realizzato e fedele al testo inserito.

Prima di mostrare e spiegare quali sono stati i prompt testuali inseriti dal candidato all'interno dei tool AI di Video Generation, è importante definire il termine Prompt Engineering.

4.7.1 Prompt Engineering

Il "prompt engineering" è un concetto fondamentale nel campo dell'intelligenza artificiale generativa, soprattutto quando si tratta di modelli di diffusione adottati per ottenere la generazione di contenuti come testi, immagini, audio e video.

Nel contesto della generazione AI di video, il prompt engineering assume un ruolo cruciale perché determina la qualità e la rilevanza del video generato in risposta a un input testuale.

Il prompt engineering consiste nell'arte e nella scienza di formulare domande o istruzioni testuali, come i prompt, in modo che un modello di diffusione di intelligenza artificiale le interpreti nel modo più efficace possibile, producendo risultati che soddisfano le aspettative dell'utente. È un processo iterativo che richiede una comprensione profonda sia del modello di diffusione AI, e dell'algoritmo

che gestisce il tool AI, sia del dominio e ambito di applicazione (generazione di video, immagini, audio, voci, testi in output).

Nel contesto della generazione di video AI, il prompt engineering è particolarmente sfidante e importante per diversi motivi:

- Complessità del contenuto: i video contengono una ricchezza di informazioni visive e auditive che devono essere coerentemente generate e sincronizzate.
- Intenzionalità: è fondamentale che il prompt catturi con precisione l'intento dell'utente, includendo stile, tono, soggetti e figure coinvolte, luoghi, azioni e qualsiasi altro dettaglio rilevante.
- Variabilità delle interpretazioni: dato che un prompt può essere interpretato in molti modi, l'ingegneria del prompt deve mirare a ridurre l'ambiguità e a guidare il modello di diffusione, usato per la generazione AI del video, verso l'interpretazione desiderata.
  Per applicare correttamente il prompt engineering, è utile seguire delle strategie per ottenere risultati ottimali e soddisfacenti. Queste sono:
- Dettaglio e chiarezza: un prompt ben progettato dovrebbe essere dettagliato e chiaro, fornendo al diffusion model tutte le informazioni necessarie per generare il video AI desiderato.
- Sperimentazione e iterazione: la creazione di prompt efficaci spesso richiede sperimentazione e iterazione, adattando il linguaggio e il livello di dettaglio in base ai risultati ottenuti.
- Personalizzazione: adattare i prompt testuali alle specifiche capacità e peculiarità del modello e del tool AI in uso può migliorare notevolmente i risultati.

L'abilità di saper generare e scrivere dei prompt testuali efficienti è anche sfidante. Infatti, il prompt engineering può offrire sfide come:

- Equilibrio tra dettaglio e flessibilità: trovare il giusto equilibrio nel fornire abbastanza dettagli senza riempire troppo la creatività del modello di diffusione.
- Gestione delle aspettative: i risultati della generazione di video possono variare significativamente in base alla complessità del prompt e alle capacità del modello, richiedendo quindi una gestione accurata delle aspettative degli utenti nei confronti dei video AI che vogliono generare.

Il prompt engineering è un elemento chiave per sfruttare al meglio le potenzialità dell'AI Generativa, specialmente nell'ambito della generazione video. Una comprensione approfondita e un'applicazione attenta delle tecniche di ingegneria del prompt possono notevolmente migliorare la qualità e la pertinenza dei video generati, rendendo questa competenza essenziale per i ricercatori e i praticanti nel campo dell'IA Generativa.

4.7.2 Prompt testuali di input

I prompt testuali di input, utilizzati all'interno degli 8 tool AI di Video Generation che sono stati analizzati nelle tabelle di uno dei paragrafi precedenti, sono stati elaborati per ottenere generazioni di video AI in output, in modo tale che questi fossero efficienti e soddisfacenti il più possibile.

La generazione video adottata è stata di tipo: text-to-video generation.

Per testare gli applicativi AI di Video Generation, sono stati scritti e definiti 3 gruppi di prompt testuali che sono molto simili tra loro ma che presentano, al loro interno, delle piccole differenze in termini di parole chiave scritte, che sono relative alle caratteristiche del tool AI utilizzato.

Gli input testuali sono stati inseriti negli 8 tool e ogni prompt di ingresso ha generato un video AI relativo di output. Dunque, per ciascun tool AI utilizzato (che in totale sono 8), sono stati usati 4 prompt testuali che sono stati trasformati e generati in 4 video differenti. Moltiplicando il numero di video ottenuti in un singolo tool (che coincide col numero di prompt testuali usati in input in ciascun applicativo) per gli 8 applicativi utilizzati, si ottiene un numero di video pari a 32 ma in realtà

complessivamente i contenuti elaborati sono stati 36. Quest'ultimo numero indica la quantità totale dei video AI generati ed ottenuti dai tool.

Una nota al calcolo complessivo è relativa al fatto che: al calcolo di 4 video ottenuti per ogni tool, moltiplicato per 8 che sono i tool totali utilizzati, bisogna aggiungere altri 4 video ottenuti con l'altro tipo di video relativo a Kaiber, in cui oltre ai 4 video Flipbook, sono stati generati i 4 video Motion. Questi ultimi fanno passare il totale dei video da 32 a 36.

I prompt testuali, usati negli 8 applicativi, sono stati scritti in maniera specifica per ciascun tool utilizzato, in modo tale che l'adattamento della scrittura dei prompt, relativi a uno specifico strumento generativo, possa portare a generazioni gradevoli.

Per quanto riguarda *RunwayML*, i prompt adottati sono:

1) "*A man walks in a park during a rainy day, there is low sunlight, there are squirrels and also cars parked on the street. High quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
2) "*A man walks in a park during a rainy day, low sunlight, small squirrel on the ground and cars parked on the street. High quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
3) "*A port with rough seas, a lighthouse emitting a bright light, boats and ships, seagulls flying in the sky. High quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
4) "*A forest during a thunderstorm with trees, logs, bushes, abandoned houses, a camper and warning road signs. High quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".

Questi 4 prompt testuali, oltre che su Runway, sono stati utilizzati anche su *Kaiber* (per le due tipologie di video generate, Flipbook e Motion, mentre la tipologia di video Transform non è stata considerata in quanto prevede solo generazione video-to-video e, quindi, non consente di inserire un testo come input per generare il video in output), su *Moonvalley* e su *Leonardo*.

Per quanto riguarda *Pika*, sono stati scritti i prompt testuali elencati di sotto, in cui, rispetto a quelli scritti sopra, presentano differenze su alcune parole scritte che sono state sottolineate:

1) "*A man walks in a park during a rainy day, there is low sunlight, there are small squirrels on the ground and cars parked on the street. 3D pixar style, high quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
2) "*A man walks in a park during a rainy day, low sunlight, small squirrel on the ground and cars parked on the street. 3D pixar style, high quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
3) "*A port with rough seas, a lighthouse emitting a bright light, boats and ships, seagulls flying in the sky, 3D animation style. High quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
4) "*A forest during a thunderstorm with trees, logs, bushes, abandoned houses, a camper and warning road signs, 3D animation style. High quality render, UHD, 8k, ray-tracing, high dynamic*

range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic".

Per l'applicativo *Genmo Replay* invece si rivelano i seguenti input testuali con le differenze, rispetto ai prompt usati per Runway, che sono rappresentate dalle parole sottolineate:

1) "*A man walks in a park during a rainy day, there is low sunlight, there are squirrels and also cars parked on the street. 3D render style, high quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
2) "*A man walks in a park during a rainy day, low sunlight, small squirrel on the ground and cars parked on the street. 3D render style, high quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
3) "*A port with rough seas, a lighthouse emitting a bright light, boats and ships, seagulls flying in the sky. 3D render style, high quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".
4) "*A forest during a thunderstorm with trees, logs, bushes, abandoned houses, a camper and warning road signs. 3D render style, high quality render, UHD, 8k, ray-tracing, high dynamic range, hyper-realistic, masterpiece, lifelike, precise, photorealism, rich colors, lifelike textures, best quality, detailed, realistic*".

I tool AI video generativi Plaiday, Neverends seguono gli stessi prompt testuali dati in input a Genmo Replay.

Come si può notare, i prompt testuali inseriti e testati negli 8 tool AI sono raccolti in 3 gruppi, che differiscono in qualche parola chiave. Infatti ci sono 3 gruppi di prompt testuali:

- Quelli usati per RunwayML, Kaiber, Moovalley, Leonardo;
- Quelli usati per Pika;
- Quelli usati per Genmo, Plaiday, Neverends.

Le differenze nei 3 gruppi di prompt testuali si notano attraverso i termini sottolineati, che sono stati scritti nei diversi prompt testuali.

I 4 prompt testuali iniziali che sono stati testati e presi come riferimento, sono quelli assegnati a RunwayML e agli altri tool AI che presentano gli stessi input testuali, mentre gli altri due gruppi di prompt, adottati negli altri tool, hanno preso ispirazione da quelli inseriti su RunwayML ma presentano alcune parole chiave diverse.

Il motivo per la quale sono state usate alcune parole diverse all'interno dei prompt testuali è dovuto alla diversa predisposizione che i tool AI di video generation possiedono.

Infatti, il primo gruppo di tool AI composto da Runway, Kaiber, Moonvalley e Leonardo presenta un elemento comune: i tool hanno, al loro interno, diversi stili visivi e grafici che possono essere scelti da un elenco di opzioni (presente nelle diverse interfacce degli applicativi fruiti da browser e da Discord) e, una volta scelto lo stile, lo si può applicare ai video da generare e, nel caso dei prompt scritti, sono stati scelti e cliccati dal candidato gli stili:

- "3D render" per i 4 prompt di RunwayML;
- "3D rendering" per gli 8 prompt di Kaiber;
- "3D animation" per i 4 prompt, rispettivamente, di Moonvalley e Leonardo.

Nel caso di Pika, invece, i prompt testuali sono stati scritti specificando, tramite pure scrittura, lo stile che si è voluto adottare. Infatti, i due stili “3D pixar style” e “3D animation style” sono stati espressi in modo scritto dal candidato, senza che questi venissero scelti da un menu o da qualche interfaccia o elenco.

Infine, per il gruppo di tool AI costituito da Genmo, Plaiday e Neverends si ha la stessa peculiarità definita con Pika: lo stile visivo definito nel prompt testuale è stato inserito e scritto tramite tastiera, e non invece scelto da un elenco di opzioni di stili grafici e visivi. Quindi, il termine “3D render style” è stato espresso dal candidato durante la scrittura dei prompt inseriti nei 3 tool.

Un ultimo aspetto da considerare è relativo alla scelta di aver scritto i primi due prompt testuali in modo identico, a meno dei termini “there is” e “there are”. Questi sono stati inseriti nel primo tipo di prompt, mentre nel secondo input testuale sono stati omessi. La motivazione, di scrivere in un prompt i termini e di rimuoverli dall’altro, è dovuta alle possibili differenze di generazione AI di video che si possono ottenere in output, quando si utilizzano o meno le due forme grammaticali. Infatti, pur essendo molto simili i due input testuali, si è notato che il prompt contenente le due espressioni ha generato, per gli 8 tool, video AI diversi rispetto a quelli ottenuti a partire dal prompt testuale scritto senza quelle parole.

Quindi, nella presenza o meno di specifiche parole all’interno dei prompt testuali, si evince come variano le generazioni dei video AI che si producono in output.

Infine, rilevanti e importanti sono stati i parametri tecnici inseriti in ciascun prompt testuale, che hanno consentito maggiore qualità, risoluzione, precisione e dettaglio alla generazione dei video AI prodotti. In questo modo, i contenuti generati sono stati realizzati in maniera efficiente e soddisfacente in termini visivi. I parametri tecnici, scritti all’interno dei prompt testuali, sono individuabili a partire dalla seconda parte di testo.

4.7.3 Video di output: generazioni AI a confronto

I 36 video di output, ottenuti dalle generazioni AI eseguite negli 8 applicativi video generativi, sono stati confrontati tra di loro e valutati in maniera soggettiva e oggettiva.

Relativamente al confronto dei video AI ottenuti dagli 8 strumenti, in questo paragrafo sono presenti una serie di immagini che mostrano alcune delle generazioni video prodotte, con la presenza dei prompt testuali scritti e utilizzati come input per ottenere i video di output.

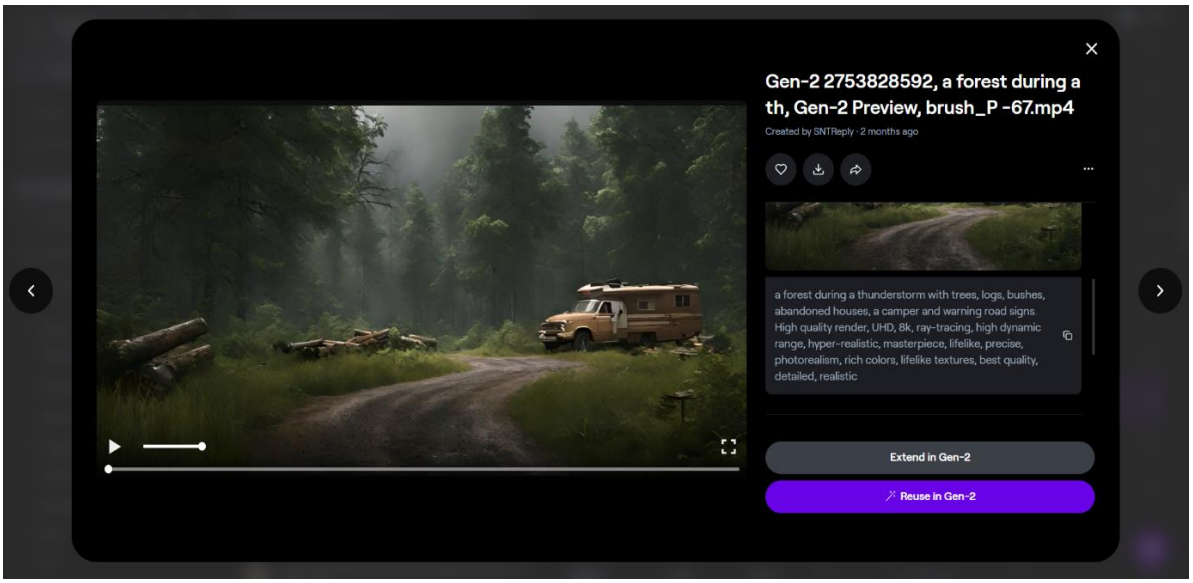

Figura 4.10 Generazione video AI ottenuta su Runway tramite prompt testuale (si tratta del prompt numero 4 dell’elenco di RunwayML scritto nel paragrafo precedente) inserito in input e mostrato alla destra della schermata

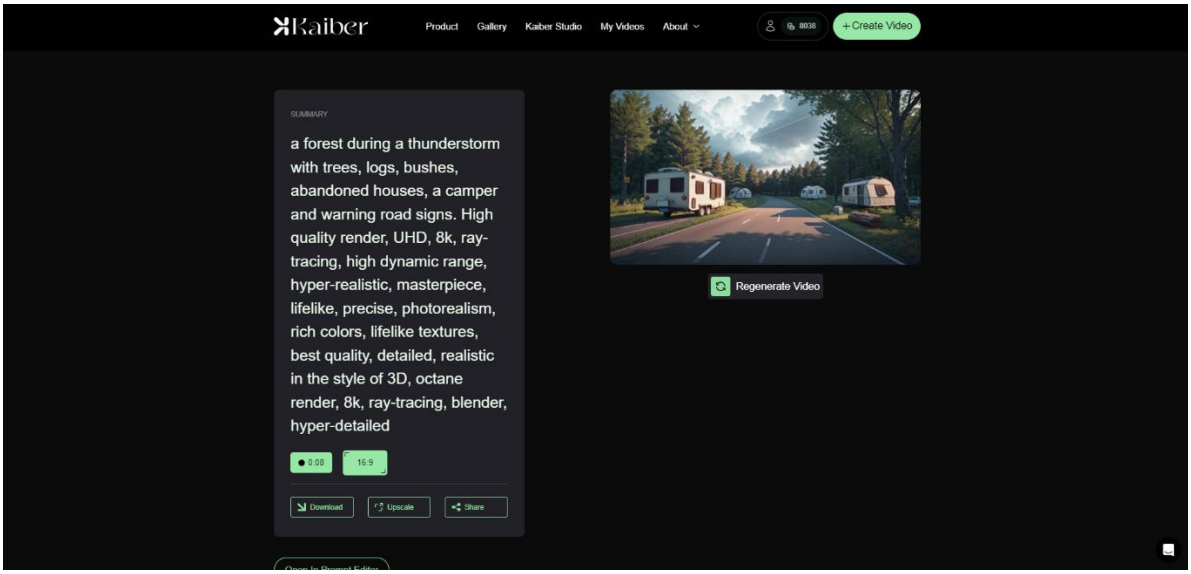

Figura 4.11 Generazione video AI (di tipo Flipbook) ottenuta su Kaiber tramite prompt testuale (si tratta del prompt numero 4 dell'elenco di RunwayML scritto nel paragrafo precedente) inserito in input e mostrato alla sinistra della schermata

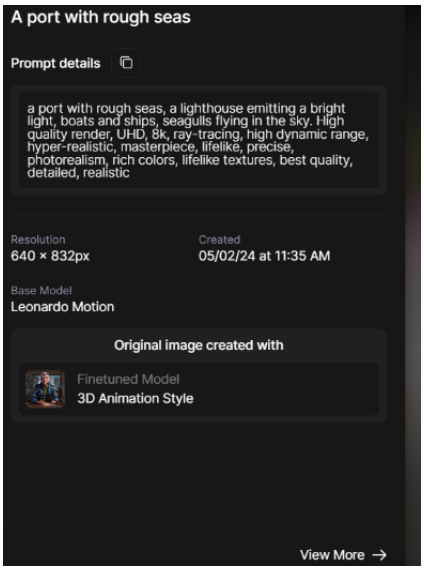

Figura 4.12 e Figura 4.13 Generazione video AI (o motion) ottenuta su Leonardo tramite prompt testuale (si tratta del prompt numero 3 dell'elenco di RunwayML scritto nel paragrafo precedente) inserito in input e mostrato nella schermata destra

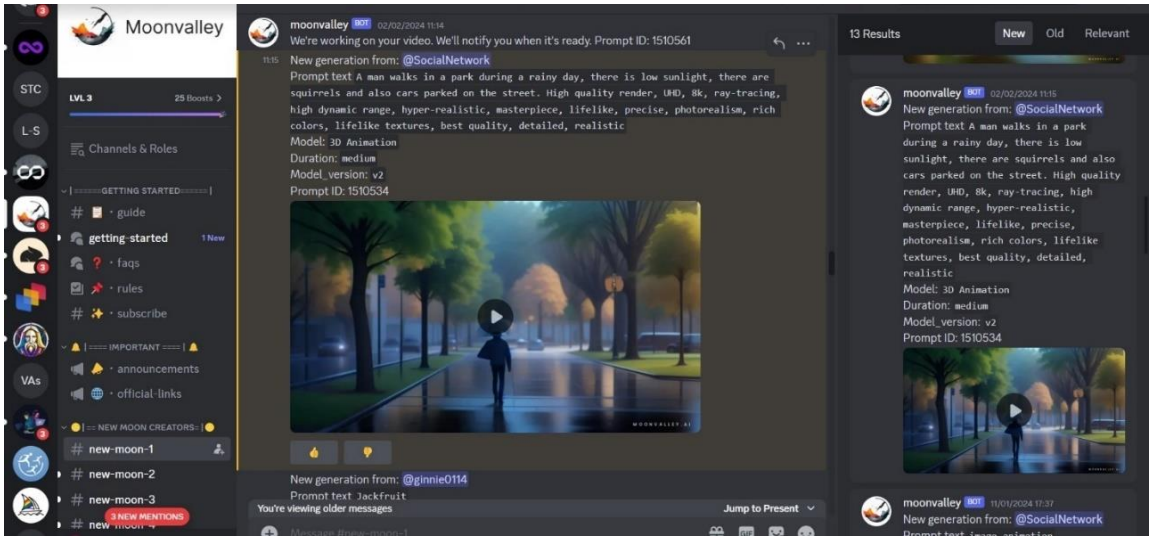

Figura 4.14 Generazione video AI ottenuta su Moonvalley tramite prompt testuale (si tratta del prompt numero 1 dell'elenco di RunwayML scritto nel paragrafo precedente) inserito in input e mostrato in alto, al centro della schermata

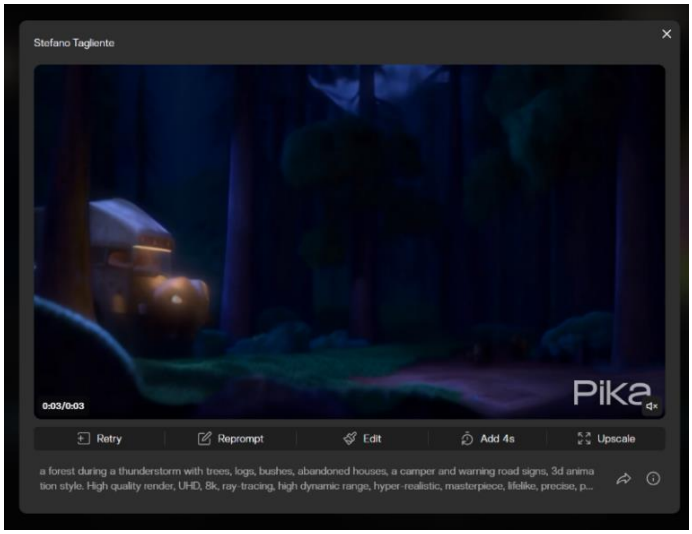

Figura 4.15 Generazione video AI ottenuta su Pika tramite prompt testuale (il numero 4 dell'elenco di Pika scritto nel paragrafo precedente) inserito in input e mostrato nella parte inferiore della schermata

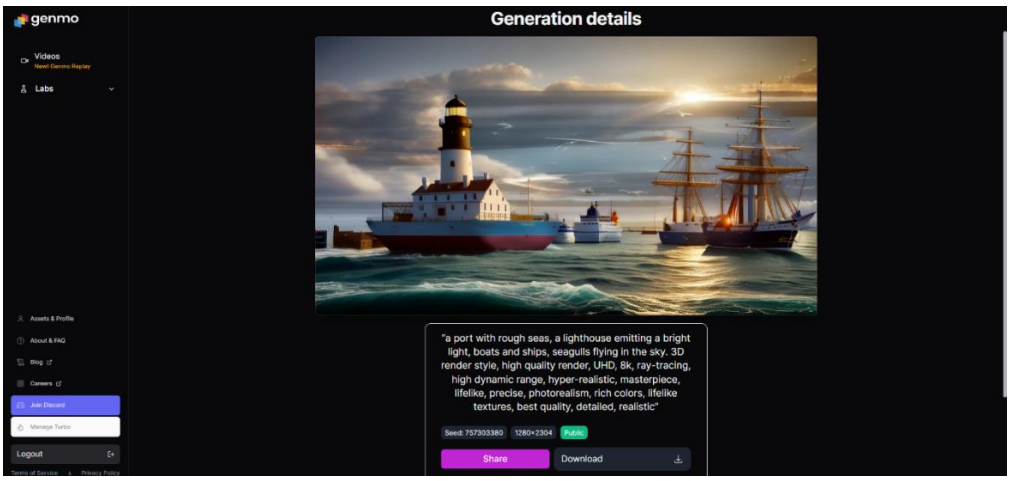

Figura 4.16 Generazione video AI ottenuta su Genmo tramite prompt testuale (il numero 3 dell'elenco di Genmo scritto nel paragrafo precedente) inserito in input e mostrato nella parte inferiore della schermata

Figura 4.17 Generazione video AI ottenuta su Plaiday tramite prompt testuale (il numero 3 dell'elenco di Genmo scritto nel paragrafo precedente) inserito in input

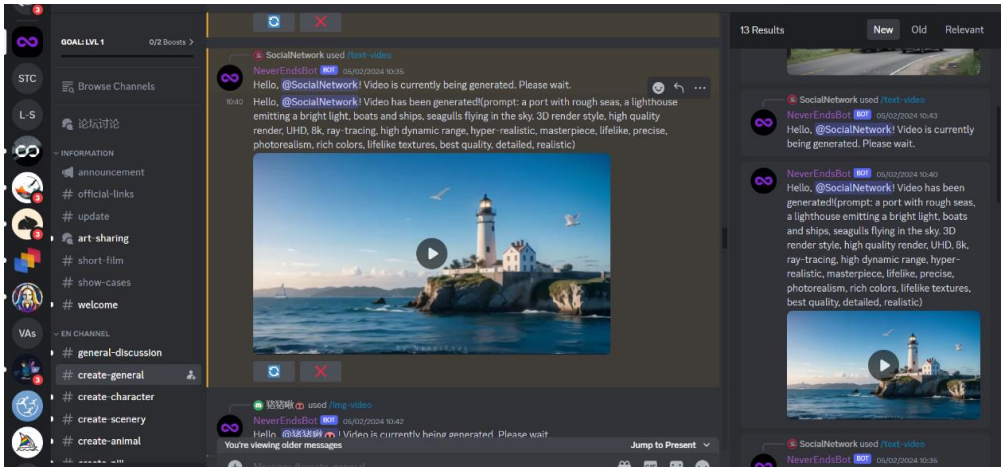

Figura 4.18 Generazione video AI ottenuta su Neverends tramite prompt testuale (il numero 3 dell'elenco di Genmo scritto nel paragrafo precedente) inserito in input e mostrato nella parte superiore della schermata

4.7.4 Overall soggettivo sulle generazioni ottenute

I 36 video AI, ottenuti dagli 8 applicativi di AI generativa, sono stati valutati, e dunque confrontati, soggettivamente dal candidato. Questo procedimento è stato utile per definire:

- Quali tool AI sono stati i migliori nelle generazioni video prodotte;
- Quali video AI sono stati definiti come migliori in termini di qualità e rilevanza.

Il primo applicativo valutato è stato *RunwayML*.

Questo applicativo, tramite i prompt testuali scritti, e arricchiti dai parametri tecnici inseriti, ha generato ottimi risultati di rendering finale dei video di output dimostrando anche una buona realizzazione degli elementi statici. Mentre alcuni elementi dinamici potevano essere gestiti meglio, in quanto bisognava ottimizzare il movimento con operazioni precise di "Motion Brush" e di modifica, eventuale e remota, del prompt. Durante l'applicazione dell'operazione di Motion Brush su uno o più elementi, si nota come viene resettato il valore di General Motion impostato nella generazione video. Inoltre, gli elementi su cui non si applica il motion brush sono stati lasciati fermi o sono stati soggetti a minimi movimenti e, dunque, a piccole animazioni.

L'overall finale sul tool AI citato, e sul video da esso generato, definisce:

- Il rendering 3D finale del video come ottimo;
- Gli elementi statici generati nel video come ben realizzati;
- Gli elementi dinamici generati nel video come buoni e, a tratti, discreti, con un uso del motion brush che poteva essere realizzato in maniera più precisa.

Proseguendo nel giudizio dei tool AI di Video Generation, che hanno generato video AI a partire dal prompt testuale di ingresso, è stato valutato *Kaiber*.

All'interno dell'applicativo sono state eseguite generazioni video per le tipologie Flipbook e Motion (che sono quelle che accettano il prompt testuale in input), mentre i video di tipo Transform non sono stati generati, in quanto non accettano il prompt testuale come input per la generazione dei video da ottenere in output. Dunque, Kaiber con i video Transform non consente una text-to-video generation, ma permette solo una video-to-video generation.

Tramite questo applicativo sono stati generati 8 video AI, 4 di tipo Flipbook e 4 di tipo Motion.

Il giudizio espresso dal candidato sui 4 video Motion definisce:

- I video come contenuti che presentano una scadente narrazione del contenuto;
- La qualità del 3D è buona, sia per gli elementi statici che per quelli dinamici;
- Gli elementi in movimento, a volte, non sono connessi tra di loro e, in certi frame, si ibridano generando animazioni anomale e movimenti innaturali.

Kaiber è un tool che, attraverso prompt testuali più semplici e corti, può generare dei buoni Motion in output, con video che mostrano una buona qualità e che seguono una logica di narrazione. Dalla valutazione però è scaturito che, con i prompt testati in questa fase di progetto tesi, l'applicativo non ha prodotto dei buoni risultati di Motion in output. Infatti, il primo frame del video è ottimo e in linea alla richiesta testuale inserita, ma appena sono stati generati gli altri frame video, si è deteriorata l'animazione con movimenti anomali.

Invece, l'overall del candidato nei confronti dei 4 Flipbook è leggermente diverso. Ne segue che:

- I video Flipbook hanno una narrazione migliore rispetto ai Motion;
- Gli elementi vengono resi dinamici con una discreta precisione nei movimenti;
- La qualità del rendering 3D è buona sia per gli elementi statici che dinamici.

I video Flipbook sarebbero più funzionali se si adottassero dei prompt testuali più corti e semplici, mentre per prompt testuali come quelli dati in input il tool ha faticato, in parte, nel generare una narrazione logica all'interno dei video, ma comunque, questa tipologia di video, ha riscontrato complessivamente un risultato migliore rispetto ai video motion.

L'osservazione finale comune, espressa per i due tipi di video generati con Kaiber, afferma che il tool presenta il vincolo di avere due tipologie di video già definite e, quindi, anche se si modifica di una certa quantità il prompt testuale inserito in input, comunque il video che viene generato in output mostra le sembianze o di un flipbook o di un motion, che non sono dei video live action ripresi con camera e, inoltre, hanno un'evoluzione rapida delle scene video generate. In Kaiber, si manifesta un effettivo vincolo nel tipo di video che si vuole ottenere, senza godere di una libertà di animazione che si desidera generare, in quanto questa o è guidata dai video Flipbook o dai Motion.

Il terzo applicativo, *Genmo*, ha riscosso altre tipologie di giudizio nei confronti dei video AI generati, che sono stati valutati del candidato.

Il giudizio soggettivo dei video AI ha dimostrato che:

- Nei 4 video ottenuti è stata prodotta una discreta qualità del rendering 3D;
- Gli elementi statici e dinamici hanno mostrato una buona qualità risolutiva;

- Le animazioni degli elementi sono state apprezzate, ma possono essere migliorate;
- I movimenti, in alcuni tratti, possono essere resi leggermente più naturali e fluidi.

In alcuni video, è stato notato un leggero effetto di motion blur, durante il movimento della camera o degli elementi presenti in scena, ma nel complesso i video si comprendono e non disturbano alla vista, con l'effetto di motion blur che è poco rilevante.

Inoltre, in specifici video si manifesta un effetto di strobing, durante il movimento della camera o degli elementi presenti in scena, ma anche in questo caso i video si capiscono e l'effetto non disturba l'utente durante la visione dei video. Dunque, l'effetto di strobing è poco rilevante nei video AI generati.

Complessivamente, il tool AI Genmo Replay, dal punto di vista del candidato, mostra buone capacità di miglioramento nella generazione dei video AI e si prepara a diventare un buono strumento alternativo a RunwayML.

Passando al tool *Moonvalley*, il giudizio complessivo dei video AI generati definisce i contenuti video come:

- Elementi visivi che possiedono una buona qualità del rendering 3D;
- Gli elementi statici e dinamici sono realizzati con una certa precisione;
- Le animazioni sono discrete e i movimenti, in generale, appaiono naturali.

Il tool presenta margini di miglioramento, alcuni elementi si muovono in modo anomalo, ma per essere fruito da server Discord è un applicativo accettabile sia in termini di qualità del 3D che in termini dell'animazione.

Un'ultima osservazione scaturita è che si può migliorare l'interpretazione del prompt testuale inserito in input, in quanto alcuni video generati sono leggermente distanti, in termini di rappresentazione del contenuto, dal prompt scritto.

Nella valutazione dell'applicativo Pika, il candidato ha definito che, a volte, lo strumento esagera negli elementi 3D che devono essere generati, rendendoli troppo simili a uno stile cartone animato. Questo, si è manifestato in output, a partire dalla scrittura in input dei prompt testuali, in cui sono stati scritti i termini "3D animation style" oppure "3D pixar style" quando si è definito lo stile del video da realizzare. Infatti, i volti delle persone ma anche gli oggetti hanno mostrato mesh, rispetto per esempio a Runway, più simili a quelle dei film d'animazione. Inoltre, il tool non ha interpretato i vari parametri tecnici scritti nei prompt testuali.

Probabilmente, una migliore resa e riproduzione degli elementi che appaiono a video può essere risolta scrivendo un altro tipo di stile nel prompt come: "3D render" che è lo stile utilizzato in Runway. Questa osservazione è stata espressa perché probabilmente il tool Pika interpreta lo stile Pixar o di animazione 3D (computer animation) come più simile a uno stile di cartone animato (come "Cars"), piuttosto che più vicino al "3D Render", che è lo stile che il candidato desiderava ottenere nei video AI prodotti (quindi stili vicini alla CGI, al 3D in stile Blender o Unreal Engine).

Nel prompt testuale numero 2 usato in Pika, il candidato ha adottato, per la generazione del video AI, la funzionalità di negative prompt (da versione browser dell'applicativo) in cui ha espresso le parole come umbrella, bags, sun, bikes che sono elementi che non dovevano essere generati nel video.

Complessivamente, con Pika si sono riscontrati questi aspetti:

- Rendering 3D del video finale generato buono, in termini di qualità degli elementi generati;
- Lo stile ottenuto non è stato però fedele a quello espresso nel prompt testuale;
- Movimenti degli elementi poco rilevanti;

- Animazioni discrete ma non definibili come buone.

Per quanto riguarda Plaiday, dai video AI ottenuti a partire dai prompt testuali, sono state evidenziati una serie di aspetti:

- La qualità del rendering 3D è risultata essere buona con elementi statici e dinamici ben realizzati;
- Le animazioni non sono state molto realistiche con movimenti a volte poco naturali;
- Alcuni video non hanno rispecchiato esattamente il prompt testuale dato in input, con la mancanza di alcuni elementi che sono stati inseriti nell'input testuale, ma che nel video AI generato in uscita, non sono stati renderizzati.

In generale, l'applicativo, essendo free, genera dei video AI accettabili, anche se in alcuni contenuti audiovisivi prodotti, degli oggetti generati sono apparsi leggermente deformati.

Relativamente a Neverends, i video generati mostrano una qualità abbastanza buona del rendering 3D, ma alcuni elementi non sono resi dinamici e rimangono statici. Inoltre, nei video AI prodotti mancano degli elementi che dovevano essere generati in quanto sono stati definiti nel prompt testuale di input, ma che nel video di output non sono stati ricreati.

Infine, i movimenti e le animazioni sono apparse leggermente "finte" in alcuni aspetti, ma essendo dei video generati da server Discord, si può affermare che questi sono risultati essere accettabili in termini di qualità e resa.

Per concludere il paragrafo, Leonardo è l'ultimo tool AI di Video Generation che è stato valutato soggettivamente dal candidato, seguendo i video AI che sono stati generati tramite una text-to-video generation.

Per generare i video si è svolto un passaggio prima verso una generazione text-to-image e, in seguito, è stata usata una generazione image-to-video, quindi nel complesso si è ottenuta una text-to-video generation ma fatta a step.

Il rendering 3D ottenuto nei video è simile a quello di Pika, quindi con uno stile cartone animato come la Pixar ed è stata notata una certa differenza tra i rendering 3D delle immagini ottenute e i rendering 3D dei video motion generati a partire dall'immagine. Infatti, l'applicativo AI si è dimostrato molto più affidabile per le generazioni AI delle immagini piuttosto che per quelle dei video. La qualità e la resa dei contenuti è molto buona, ma dipende da cosa si vuole ottenere e, quindi, dal contenuto che sono stati inseriti nel prompt testuale.

Definiti questi giudizi soggettivi riguardanti i 36 video AI (ottenuti dalle diverse generazioni text-to-video effettuate a partire dai prompt testuali), ed espresse le relative e conseguenti valutazioni sugli 8 tool AI video generativi adottati, è stata espressa una classifica soggettiva da parte del candidato che, dall'alto verso il basso, ha classificato gli applicativi utilizzati e testati.

Quindi, partendo dalla parte superiore fino a raggiungere le ultime posizioni, si trovano i tool AI video generativi che sono stati più efficienti e soddisfacenti nelle generazioni text-to-video prodotte, fino a raggiungere le parti basse con gli applicativi AI che sono stati meno abili nella creazione e generazione dei video AI. Osservando la classifica stilata, si evince, anche, come i video dei rispettivi tool AI classificati, sono posizionati allo stesso modo.

La classifica finale stilata dal candidato e relativa agli 8 tool AI di tipo Text-to-Video Generation utilizzati, è mostrata di seguito:

1. RunwayML
2. Genmo
3. Kaiber
4. Leonardo
5. Neverends
6. Plaiday
7. Moonvalley
8. Pika

Per aver stilato questa classifica, il candidato ha considerato diversi parametri di valutazione soggettiva. I parametri rispettati sono:

- Qualità del rendering e della risoluzione;
- Gli effetti di luci e ombre;
- Il 3D generato a partire dal testo;
- Il fotorealismo prodotto;
- La gamma di colori adottata;
- La precisione dei dettagli e delle textures sugli oggetti generati.

4.7.5 Pareri esterni dal panel

Dal form compilato da un numero di utenti, sono stati evidenziati dei risultati relativi alle valutazioni e ai giudizi espressi in merito alle generazioni Text-to-Video, ottenute dagli 8 applicativi AI video generativi utilizzati.

I parametri seguiti e rispettati per le valutazioni oggettive date dagli utenti esterni sono:

- Fedeltà del video rispetto al prompt testuale inserito in input;
- Precisione e accuratezza degli oggetti generati;
- Quantità degli oggetti generati;

Di seguito, sono mostrate delle schermate contenenti alcuni risultati ottenuti dalla compilazione del form o panel di valutazione. Il form, ideato e realizzato dal candidato, è stato composto in 3 sezioni, ognuna riguardante un parametro di valutazione.

Figura 4.19 Spiegazione del form compilato dal panel che precede l'inizio della sezione 1 relativa alla fedeltà dei video AI rispetto al prompt testuale inserito in input

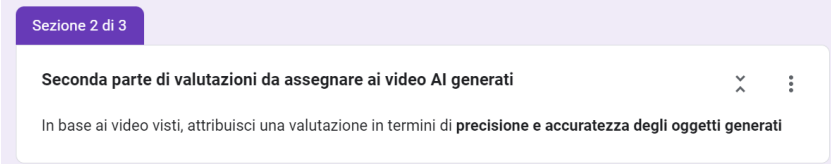

Figura 4.20 Inizio sezione 2 del form votato dal panel

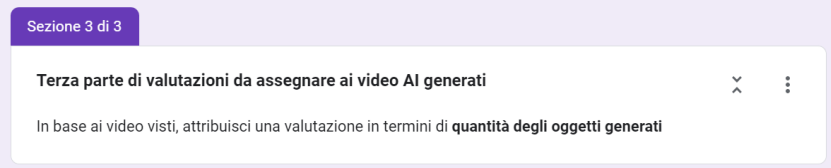

Figura 4.21 Inizio sezione 3 del form votato dal panel

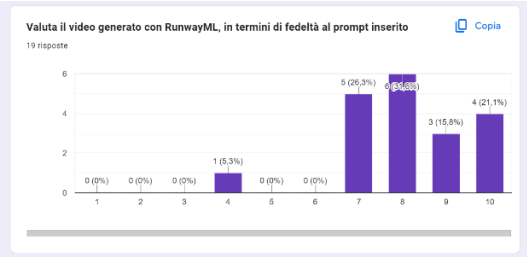

Figura 4.22 Valutazione di Runway sez. 1

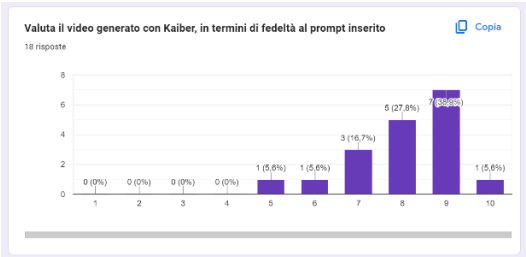

Figura 4.23 Valutazione di Kaiber sez. 1

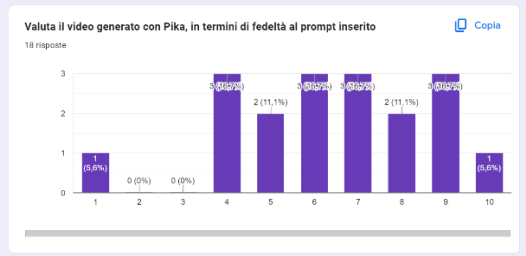

Figura 4.24 Valutazione piuttosto negativa di Pika sez. 1

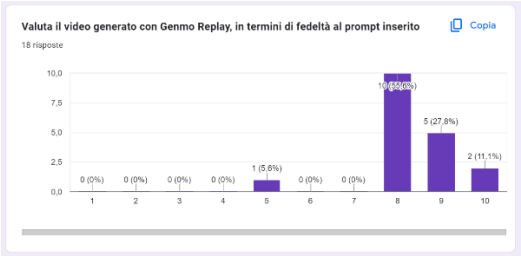

Figura 4.25 Valutazione di Genmo sez. 1

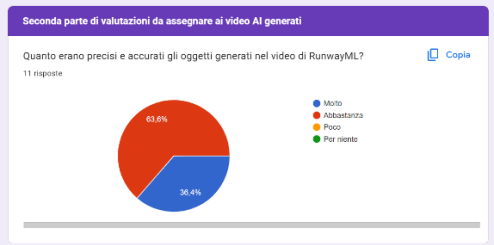

Figura 4.26 Inizio sez. 2 valutazioni positive assegnate a Runway

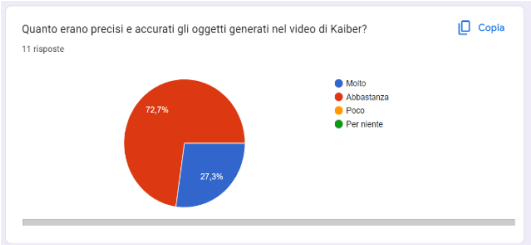

Figura 4.27 Valutazioni di Kaiber sez. 2

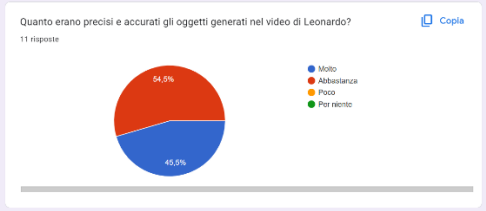

Figura 4.28 Valutazioni di Leonardo sez. 2

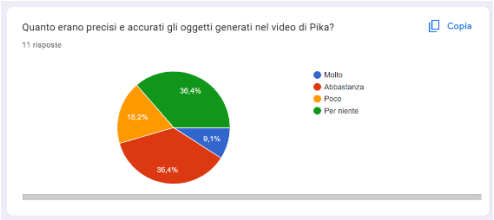

Figura 2.29 Valutazioni piuttosto negative di Pika sez. 2

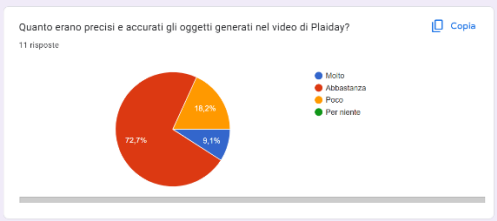

Figura 2.30 Valutazioni di Plaiday (discord) sez. 2

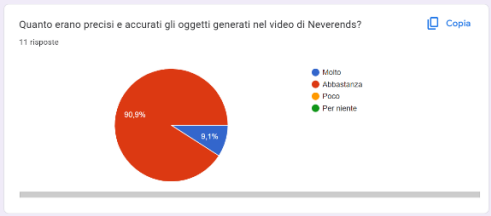

Figura 2.31 Valutazioni ottime di Neverends (discord) sez.2

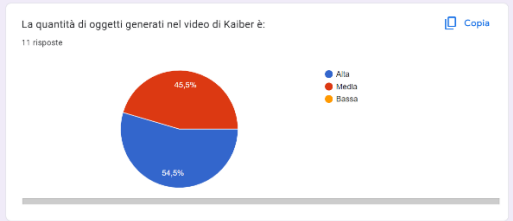

Figura 2.32 Valutazioni di Kaiber sez. 3

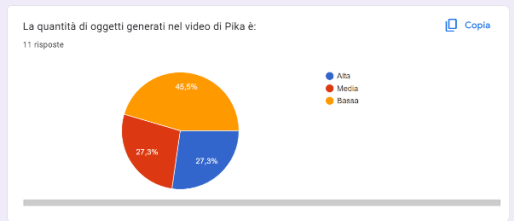

Figura 2.33 Valutazioni negative di Pika sez. 3

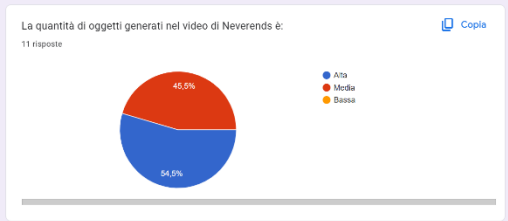

Figura 4.34 Valutazioni ottime di Neverends (discord) sez. 3

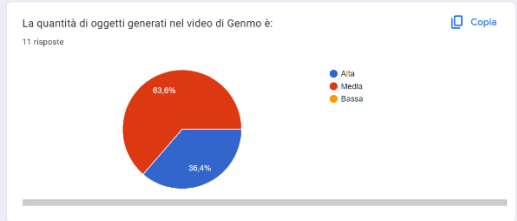

Figura 4.35 Valutazioni buone di Genmo sez. 3

Dalle risposte ricevute dagli utenti, si è notato come alcuni tool AI come Runway, Kaiber e Genmo hanno rispettato le aspettative nelle 3 sezioni del form, mentre invece l'applicativo Pika, per le generazione di video AI ottenuta, si è dimostrato poco efficiente relativamente ai 3 parametri scelti per condurre il form di valutazione.

Infine, due dei tre strumenti AI video generativi fruibili da Discord come Plaiday e Neverends hanno avuto, nelle 3 sezioni, un buon apprezzamento da parte dei votanti al form creato.

Le valutazioni espresse dal panel sono relative alle generazioni text-to-video ottenute dagli 8 tool AI Video Generativi. Le valutazioni, date per ogni tool, dipendono strettamente dagli specifici prompt testuali inseriti in input dal candidato durante questa parte di confronto e test.

I giudizi non sono, dunque, definibili come universali per qualsiasi prompt testuale che viene inserito in ingresso.

Quindi, nel caso di Pika, si ha avuto un riscontro piuttosto negativo nelle 3 sezioni e nell'ottica dei prompt testuali specifici, scelti e scritti in input dal candidato. Questo, non testimonia il fatto che Pika è però considerato un pessimo tool in generale in quanto, per altri e diversi prompt testuali inseribili in input, l'applicativo potrebbe offrire buoni risultati.

5. Produzione e realizzazione serie video AI

5.1 Obiettivo di realizzazione e target

L'altra parte del progetto di tesi svolto dal candidato è relativa alla produzione e realizzazione di una serie originale di video AI, prodotti durante il percorso collaborativo con l'azienda Reply.

All'interno di questo paragrafo sono definiti due elementi fondamentali che sono stati utili per la produzione e la realizzazione dei video AI, relativi a dei progetti aziendali svolti in passato da Reply e scelti, a piacimento, dal candidato.

L'obiettivo di realizzazione della serie originale di video AI consiste in una completa dimostrazione delle opportunità e possibilità che la Generative AI offre per la generazione e creazione di contenuti multimediali. Infatti, i lavori realizzati mostrano come, tramite un utilizzo accurato e preciso dell'Intelligenza Artificiale generativa, si possono produrre dei risultati soddisfacenti che seguono uno specifico flusso di lavoro, ideato e applicato dal candidato.

I video AI realizzati hanno l'obiettivo di intrattenere l'utente e di coinvolgerlo nella completa narrazione e visione, guidandolo in un percorso audiovisivo realizzato interamente con l'ausilio dell'AI Generativa. Infatti, i prodotti realizzati mirano a mantenere alta la visione e l'attenzione dell'utente, con l'obiettivo che quest'ultimo rimanga attivo lungo tutta la durata dei video.

La Generative AI è stata trattata in tutte le sue parti: da quella relativa a Video e Image Generation, a quella legata a Text e Voice Generation.

Questo lavoro di produzione e realizzazione di video AI è la testimonianza del fatto che la nuova AI Generativa è un utile strumento per creare contenuti multimediali in modo efficiente e piuttosto soddisfacente. Inoltre, dimostra come i tool AI Video Generativi possono essere un ottimo supporto ai software di video editing e di post-produzione, come Adobe Premiere, Adobe After Effects e DaVinci Resolve, in quanto gli applicativi AI, tramite lo svolgimento di un certo numero di operazioni, offrono la possibilità di creare diverse clip video piuttosto precise ed elaborate, senza utilizzare, eventualmente, in modo completo ed approfondito i citati programmi di editing.

Nella maggior parte dei casi, come accaduto nel progetto svolto dal candidato, i risultati di output, ottenuti con l'AI Video Generativa, vengono uniti all'interno di un software di video editing, senza però apportare grosse modifiche a questi. L'unione delle sequenze video AI è importante in quanto porta alla creazione del video finale, composto dalle diverse scene video generate dall'intelligenza artificiale.

Dunque, un altro obiettivo del progetto di tesi condotto è dimostrare che la produzione multimediale, tramite l'uso dell'AI Generativa, può cambiare in maniera notevole, innovando nelle tecniche e nelle generazioni da applicare e portando all'utilizzo di nuovi strumenti più facili ed intuitivi.

Relativamente al target di riferimento coinvolto, questo è riconducibile ai dipendenti dell'azienda Reply. Infatti, i video AI realizzati, relativi ai progetti aziendali, sono stati pubblicati sulla piattaforma video interna all'azienda, a cui hanno accesso i soli dipendenti. La fruizione e la visione completa dei video è, dunque, consentita ai dipendenti e al personale di Reply.

5.2 Workflow seguito per la realizzazione

Per la produzione e realizzazione dei video AI relativi ai progetti aziendali, è stato applicato uno specifico workflow o flusso di lavoro. Questo è stato utile per organizzare il lavoro da compiere e per svolgere le diverse fasi in modo preciso e ordinato.

Si specifica che il flusso di lavoro adottato dal candidato è personale e, dunque, ogni altro creatore di contenuti e utente può seguirne uno proprio, che può essere totalmente diverso oppure simile a quello esplicato in questo paragrafo.

Il flusso di lavoro eseguito consiste in una serie di fasi.

Prima di trattare le fasi svolte nel dettaglio, è utile specificare che i diversi step, rispettati per la creazione dei video AI aziendali, sono stati distribuiti in 3 macro-tipologie di fasi:

1. Una fase di pre-produzione in cui il candidato ha elaborato, a partire dalle schede dei progetti aziendali considerati, degli script testuali che raccontano e narrano, nella loro interezza, i progetti svolti da Reply. Gli script testuali creati, sono stati inseriti all'interno di diverse scene video e queste sono state smistate in diverse sezioni, per avere un'organizzazione migliore in termini di narrazione e flusso del video.
   Le sezioni, pensate e adottate per una precisa distribuzione delle scene video, sono: *Introduzione*, *Sfide* (*Challenges*), *Soluzioni* (*Solutions*) e *Conclusione*.
2. Una fase di produzione in cui il candidato ha generato delle immagini, prima di storyboard e successivamente definitive, per i video realizzati. Le immagini definitive rappresentano i diversi frame video che compongono il prodotto audiovisivo finale. Oltre alle immagini, sono state generate le animazioni di queste e, perciò, sono state prodotte le diverse clip o scene video.
3. Una fase di post-produzione in cui il candidato ha unito le clip video, generate dai tool AI Video Generativi, all'interno di un software di video editing. Questo è stato fondamentale per ottenere il video completo con tutte le sequenze video inserite in serie. Inoltre, è stata svolta una fase di generazione vocale degli script, che consiste nella loro trasformazione in voci narranti che fungono da voice over nei video e che narrano il contenuto generato.

5.2.1 Script e narrazione: chatGPT

La prima fase eseguita nel flusso di lavoro adottato dal candidato, è relativa alla generazione di script testuali, utili alla narrazione dei video AI realizzati. I contenuti audiovisivi generati riguardano dei progetti aziendali.

Come citato nel paragrafo 5.2, gli script testuali prodotti sono stati inseriti all'interno di diverse scene video, per organizzare meglio il racconto e la narrazione dei progetti aziendali descritti nei video.

Le diverse scene video, contenenti i relativi script testuali elaborati, sono state distribuite in 4 sezioni, interne ai video, che sono:

- Sezione di *Introduzione* del progetto, contenente scene video introduttive che spiegano lo scenario di partenza e il contesto in cui l'azienda ha dovuto lavorare;
- Sezione relativa alle *Sfide* affrontate dall'azienda durante lo sviluppo del progetto aziendale (svolto per determinati clienti);
- Sezione relativa alle *Soluzioni* applicate da Reply per risolvere le differenti tasks di progetto e per trasformare le sfide in risultati concreti;
- Sezione finale con le *Conclusioni* relative al progetto aziendale, realizzato da Reply e narrato e trattato nel video relativo.

Queste sezioni sono state considerate dal candidato in quanto, nella scheda di progetto dei progetti aziendali di Reply, sono presenti le stesse denominazioni di sezioni, contenenti le diverse informazioni.

Inoltre, rilevante è l'aspetto relativo alla lettura delle schede tecniche dei progetti aziendali scelti, dalla quale il candidato ha estrapolato le informazioni principali per elaborare gli script dei video.

Per la fase di generazione degli script, narrazione dei video AI e suddivisione del progetto in scene video, è stata considerata, come punto di partenza, la scheda tecnica dei progetti aziendali, realizzata dai team aderenti al progetto.

Il tool AI di Text Generation utilizzato dal candidato, per la generazione degli script e la distribuzione di questi in differenti scene video e sezioni, è ChatGPT (nello specifico è stata utilizzata la versione 4 dell'applicativo sviluppato da OpenAI).

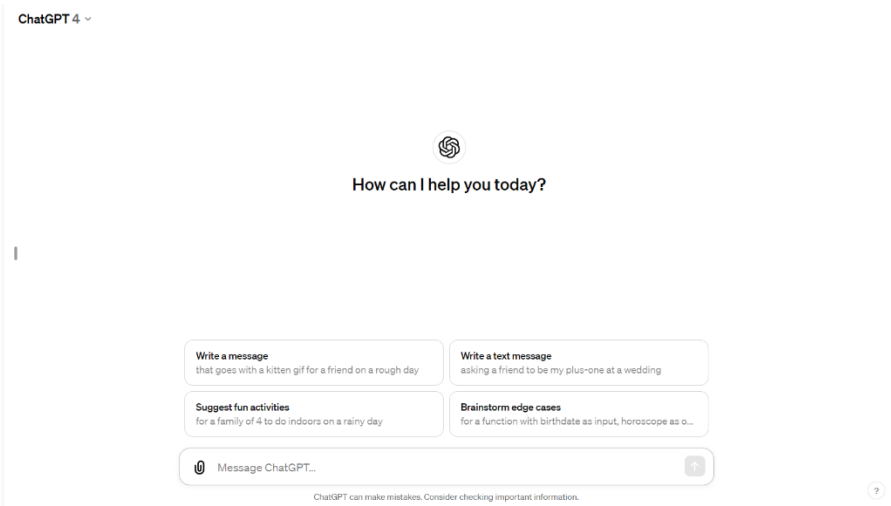

Figura 5.1 Homepage del tool AI ChatGPT (versione 4 selezionata e visibile in alto a sinistra)

Gli script sono stati generati ed elaborati tramite diverse versioni di approcci, adottati e creati dal candidato. Dopo un'attenta valutazione, delle tre versioni di approcci utilizzati, ne è stato selezionato uno come definitivo e adeguato.

I 3 approcci, spiegati in maniera approfondita, sono stati testati sui diversi progetti aziendali scelti dal candidato ed è stato valutato e scelto l'approccio migliore per la generazione degli script testuali, raccolti nelle scene video che narrano i progetti.

Il primo approccio, testato nella prima fase del workflow, consiste nell'inserimento delle sezioni della scheda tecnica, relativa al progetto aziendale scelto, all'interno di chatGPT 4, a cui è stato aggiunto un vincolo sulla durata complessiva che si desidera attribuire al video da generare. Tramite questo approccio sono state generate diverse versioni di script, tra cui quelli intermedi che si sono rivelati essere i più efficaci.

L'approccio 1, nello specifico, consiste nell'inserimento in input a chatGPT 4, della scheda tecnica del progetto aziendale che si è deciso di trattare, in cui sono state inserite le parti di Introduzione, di Sfide o Challenges e di Soluzioni o Solutions.

Inserite queste 3 sezioni, è stato chiesto al tool AI di generazione testuale, di generare una suddivisione in scene video del progetto, con relativi script testuali narranti. Successivamente, è stato comandato all'applicativo di generare una scena video per la parte di conclusione con uno script adeguato.

Fornite in input le parti di testo delle 3 sezioni (introduzione, sfide, soluzioni), il candidato ha eseguito la seguente richiesta al tool: effettuare una generazione di scene video, con relativi script adatti per ogni porzione di testo (sezione) passata in input.

In seguito, una volta che il tool ha generato la prima versione di scene video con gli script relativi, è stato impostato il vincolo di lunghezza del video, per ottenere una durata degli script più breve.

Quindi, dopo la prima generazione di scene video e script relativi alle sezioni di Introduzione, Sfide e Soluzioni, è stato richiesta, all'applicativo, la generazione di script più brevi (che avessero una lunghezza più ridotta) e, in seguito, quella di script intermedi (che avessero una lunghezza intermedia).

Le scene video con gli script intermedi si sono rivelate essere le migliori, rispetto a quelle ottenute con la prima generazione di script iniziali (e lunghi) e rispetto a quelle con gli script più brevi ottenuti nella seconda generazione. La preferenza è stata eseguita in termini di contenuto e narrazione del progetto ma anche in termini di lunghezza e durata complessiva che si è attribuita al video relativo al progetto aziendale.

Dunque, ottenuti gli script lunghi, brevi e intermedi, sono stati selezionati gli script intermedi con le relative scene video generate per le 3 sezioni di progetto: introduzione, sfide, soluzioni.

Dopo aver terminato le 3 sezioni del video relativo al progetto scelto, con la creazione di diverse scene video e differenti script generati, è stato chiesto al tool di creare una scena finale di conclusione, che potesse concludere il video attraverso uno script adatto.

Anche allo script della scena finale, è stato utilizzato il vincolo di rispettare una certa lunghezza e di essere, perciò, breve. Infatti, la prima generazione della scena finale si è dimostrata essere troppo lunga, mentre successivamente, dopo aver comando l'applicativo, è stata ottenuta una scena finale, con relativo script, di durata più ridotta e ottimale per il video da realizzare.

La durata stimata del video, assegnata al tool attraverso un prompt, era quella di essere compresa tra 1 minuto e 30 secondi e i 2 minuti e, in base a questa durata assegnata, gli script intermedi delle varie sezioni si sono dimostrati essere adeguati.

Il vincolo della durata totale del video (1:30 minuti/2 minuti) è stato inserito, nello specifico, dopo aver ottenuto la prima generazione di scene e script video delle 3 fasi di progetto, in quanto i risultati ottenuti sono sembrati troppo lunghi. Dunque, è stato utile inserire questa richiesta della durata, per ottenere degli script più brevi ma comunque affidabili per le 3 sezioni (prima sono stati generati gli script più brevi e in seguito quelli intermedi).

Una volta effettuate le generazioni complessive relative alle scene video con gli script intermedi, è stata richiesta, all'applicativo, la durata stimata del video contenente gli script intermedi. Questa richiesta è stata svolta per avere un'idea sull'ipotetica durata complessiva del video finale, ma anche per capire se la durata del video fosse in linea con quella citata in precedenza (1:30min/2min).

Questo approccio è stato provato per differenti progetti aziendali selezionati dal candidato. Oltre a questo approccio 1, sono state pensate e testate altre 2 metodologie o approcci.

Il secondo approccio della fase 1 del workflow è relativo alle parole chiave, individuate dalle sezioni della scheda di progetto aziendale scelto, unite all'assegnazione della durata di ogni sezione.

Per suddividere il progetto aziendale in diverse scene video con relativi script, è stato applicato questo approccio consistente nell'individuazione delle parole chiave presenti nelle diverse sezioni della scheda di progetto. In seguito, le keywords sono state scritte in input nel tool di OpenAI per ottenere il risultato sperato.

Assunte le diverse sezioni della scheda di progetto (introduzione, sfide, soluzioni), sono state inserite in input, a chatGPT 4, le parole chiave, raccolte dal candidato per sezioni e separate da un ";".

Il primo step è iniziato dalla sezione "Introduzione" con l'elenco delle keywords da cui generare diverse scene video con annessi script, successivamente è stata approcciata la parte di "Sfide" con la medesima logica di generazione e, infine, è stata trattata la sezione delle "Soluzioni".

Allo strumento AI, è stato specificato che ciascuna sezione del video e del progetto (introduzione, sfide, soluzioni) dovesse durare dai 20 ai 40/50 secondi, con la durata variabile in base alla sezione trattata e, attraverso questa richiesta, il tool ha generato script per le scene video che hanno mostrato una lunghezza adatta alla durata di ogni sezione video.

Successivamente alle 3 sezioni, è stato chiesto, in input, di generare una scena finale del progetto e del video, con uno script breve che non superasse i 10/15/20 secondi di durata.

Per la sezione introduttiva è stata assegnata una durata variabile tra i 20 e i 30 secondi per progetto, per la sezione “Sfide” è stata inserita una durata di 30/40 secondi, per la sezione “Soluzioni” è stata decisa una durata di 40/50 secondi e, per la parte conclusiva, è stata attribuita una durata di 10/15/20 secondi.

Un dettaglio da non sottovalutare è quello relativo al fatto che, in seguito alla prima generazione di scene video e script relativi alle parte di Introduzione, è stato notato un risultato troppo lungo in termini di script.

Notata la lunghezza elevata degli script dell’introduzione, è stata assunta la decisione di richiedere, nella successiva generazione, una sezione introduttiva che non superasse i 30/40 secondi e, quindi, le scene video e gli script relativi hanno rispettato questo vincolo temporale, diventando più brevi rispetto a quelli creati nella prima generazione.

Impostato questo vincolo temporale, le generazioni successive delle 2 sezioni rimanenti (Sfide e Soluzioni), contenenti determinate scene video e script correlati, sono state adattate automaticamente dal tool alla durata di 30/40/50 secondi.

Invece, per la sezione finale è stata chiesta una generazione piuttosto breve di scena video e script relativo e l’applicativo ha interpretato il termine "breve" come una durata di circa 10/15 secondi, da assegnare alla scena video conclusiva. Questo significa che lo script relativo alla scena video finale è dovuto rientrare in questa durata prefissata.

Anche questo approccio è stato testato, per i progetti aziendali selezionati dal candidato.

Oltre a questo secondo approccio, ne è stato ideato e provato un terzo, che conclude le metodologie adottate per comporre la fase 1 del flusso di lavoro, seguito per la realizzazione dei video AI relativi ai progetti aziendali.

Il terzo approccio della fase 1 del flusso di lavoro è collegato ai paragrafi delle sezioni della scheda di progetto aziendale, che sono stati, prima, tradotti in italiano (da inglese) e, in seguito, sono state ottenute delle versioni più brevi ma significative dei paragrafi tradotti, tramite l’assegnazione al tool di produrre i diversi concetti dei paragrafi, adottando un tono intrigante e coinvolgente.

Questo approccio, per elaborazione e fasi svolte, si è dimostrato essere il più segmentato e lungo.

A partire dalla scheda tecnica del progetto aziendale, sono stati presi separatamente i diversi paragrafi delle sezioni di Introduzione, Sfide e Soluzioni.

Estrapolate, progressivamente, le parti di testo delle 3 sezioni, è stata eseguita una traduzione del testo, da inglese a italiano, per facilitare il lavoro.

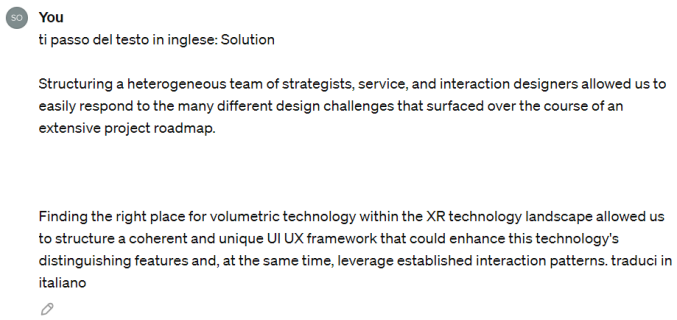

Figura 5.2 Paragrafo "Solution" estratto dalla scheda di progetto, a cui si chiede una traduzione in italiano a partire dall'inglese

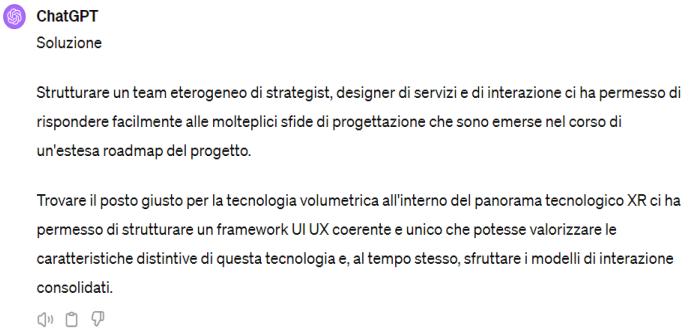

Figura 5.3 Traduzione del paragrafo "Solution" da inglese a italiano

Dai pezzi di testo tradotti in italiano, si è chiesto all'applicativo AI di effettuare una sintesi dei contenuti testuali, ottenuti in italiano, che potessero contenere i concetti principali. Inoltre, è stato richiesto di assegnare un tono intrigante, adatto per dare al video un racconto più coinvolgente.

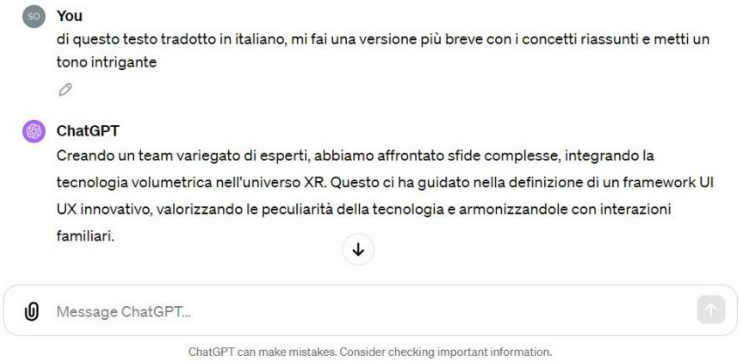

Figura 5.4 Richiesta di sintesi del testo con mantenimento dei concetti principali e inserimento di un tono intrigante e coinvolgente al contenuto testuale generato

Osservando i risultati ottenuti, si è notato come le sezioni di testo coesistono tra di loro e, una volta che si sono generate separatamente, vengono unite per formare un testo unico composto da 3 parti: introduzione sintetizzata, sfide sintetizzate, soluzioni sintetizzate.

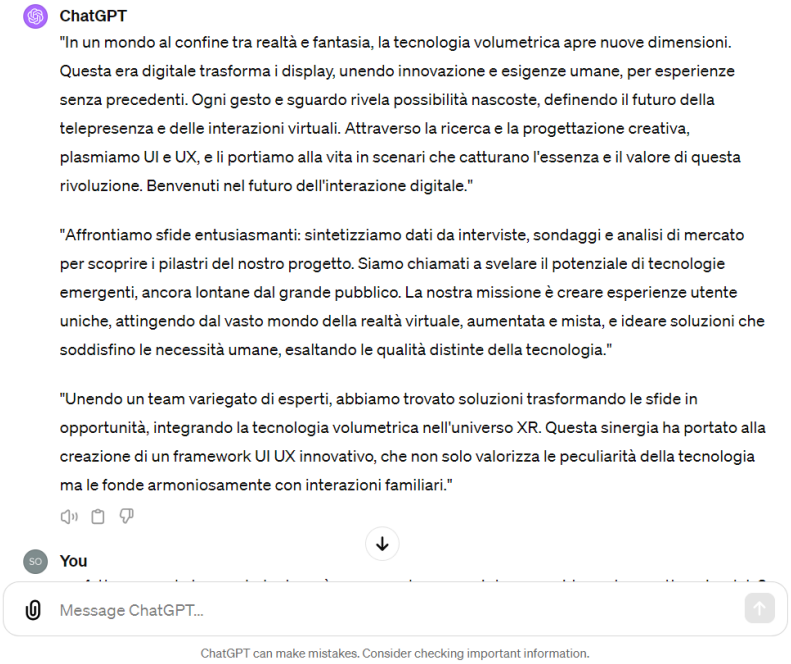

Figura 5.5 Primo testo sintetizzato ottenuto in italiano con le 3 sezioni

Al testo unico generato, si è comandato, al tool di text generation, di generare una conclusione, che in poche righe, riassumesse il progetto aziendale raccontato. Questa conclusione viene unita al testo, ottenendo, dunque, un unico discorso narrante formato da 4 parti: introduzione; sfide; soluzioni; conclusione.

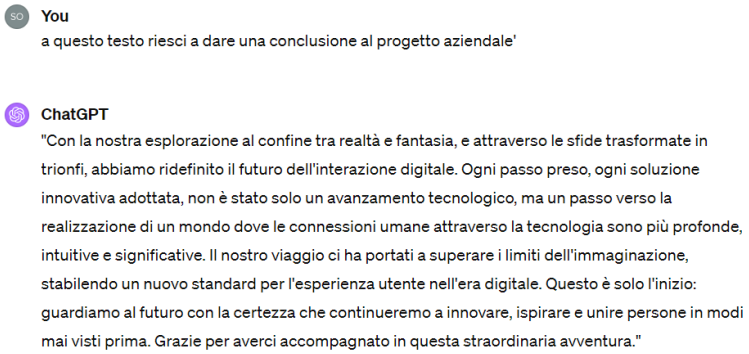

Figura 5.6 Richiesta di generazione della parte di conclusione del testo (versione più lunga della generazione)

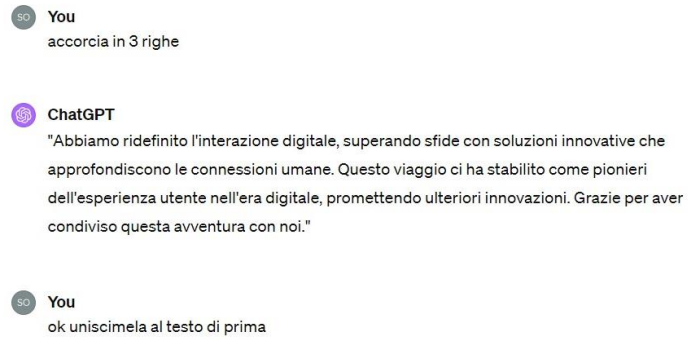

Figura 5.7 Richiesta di generare una versione ridotta della conclusione creata e mostrata nella schermata precedente, più richiesta al tool di aggregare la parte conclusiva ridotta al testo generato con le 3 sezioni

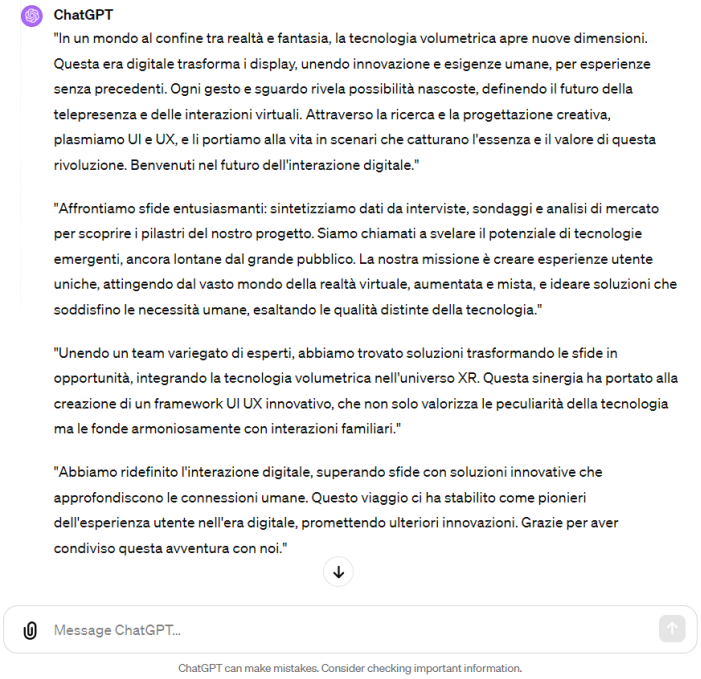

Figura 5.8 Versione testo sintetizzato con 4 sezioni ridotte ma significative: Introduzione, Sfide, Soluzioni, Conclusione

Tutte e 4 le sezioni di testo sintetizzato, sono sintetiche e contengono i concetti principali espressi con un tono intrigante.

Generato il nuovo testo o discorso del video, si è effettuata la suddivisione in scene video di questo, con una serie di script, relativi ad ogni scena video.

La scomposizione del testo in diverse scene video, con script relativi, è stata fatta prima senza i dettagli del progetto e, successivamente, è stato chiesto al tool di inserire i dettagli del progetto in alcuni script, per fornire maggiore chiarezza e assegnare uno specifico contesto al progetto aziendale da narrare nel video. Dall'applicazione di questi passaggi, è stata svolta la traduzione in inglese degli script contenenti i dettagli, in moto tale da facilitare le operazioni di generazione voce effettuate nella fase 4 del workflow, in cui il voice over è in lingua inglese.

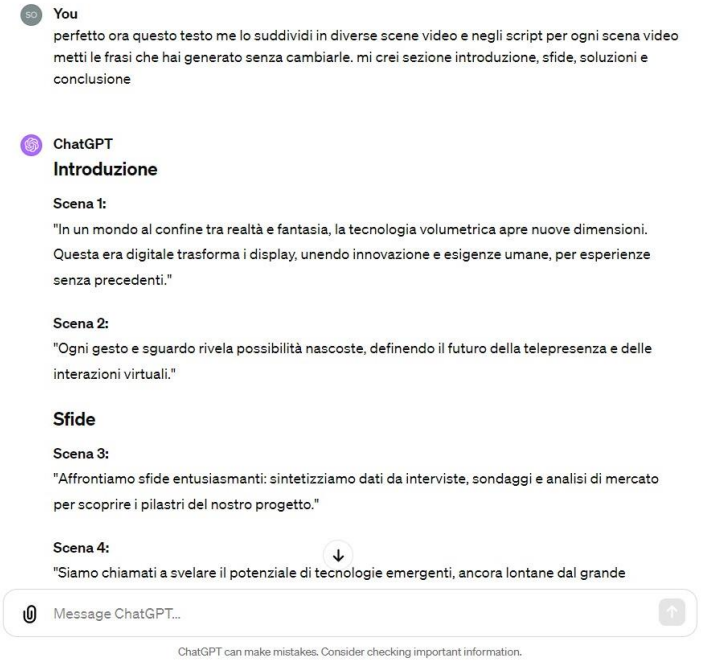

Figura 5.9 Richiesta di ottenere le scene video del progetto aziendale, con relativi script, suddivise per sezioni (nella schermata si nota l'assenza dei dettagli del progetto aziendale; la schermata con i dettagli è analoga ma presenta la richiesta al tool di inserire una serie di dettagli citati dal candidato)

Da sottolineare che, ad alcune frasi generate e usate per creare il testo unico, si è voluto assegnare il vincolo che, una volta unite, queste rientrassero in 30 secondi di durata per quella sezione di testo a cui appartengono.

Quindi, create e unite delle frasi indicanti gli script, queste sono state accorpate per creare un corpo di testo, indicante una specifica porzione di testo sintetico ottenuto e che risulta relativo al video del progetto aziendale preso in considerazione.

Spiegati i 3 approcci, un aspetto da enfatizzare, e comune alle tre metodologie testate, è quello relativo ai dettagli di progetto che sono stati inseriti negli script relativi alle scene video dei diversi progetti aziendali. I dettagli aggiunti sono:

- Nome dell'azienda che ha realizzato il progetto (che sarebbe Reply);
- La durata di realizzazione del progetto;
- La size del progetto, dunque la quantità di persone coinvolte all'interno del progetto;
- Il paese del cliente coinvolto nel progetto, e dunque il paese dell'azienda cliente per cui Reply ha lavorato;
- L'anno o gli anni di realizzazione del progetto.

Applicati i 3 approcci di generazione script sui diversi progetti aziendali, è stato notato dal candidato, che i risultati migliori (gli script generati e la suddivisione dei progetti in scene video) sono stati ottenuti tramite l'utilizzo dell'approccio 3.

Infatti, questo approccio si è dimostrato efficiente ed ottimale per generare script adatti per i video che si vogliono realizzare. La scelta dell'approccio migliore è stata fatta tramite una valutazione di diversi parametri quali:

- La lunghezza complessiva degli script e dei discorsi della voce narrante;
- La comprensione del contenuto raccontato;

- Il coinvolgimento che l'utente potrebbe avere all'ascolto degli script;
- La generazione degli storyboard che si possono ottenere a partire dagli script generati per le diverse scene video.

Dunque, gli approcci che sono stati esclusi dal workflow per la realizzazione dei video AI aziendali sono l'approccio 1 e l'approccio 2, mentre, dalla lettura degli script generati per i diversi progetti aziendali, è stato riscontrato che l'approccio 3 si è rivelato essere il migliore ed è, quindi, stato scelto.

Una nota rilevante, riguardante questa fase di workflow, è relativa a un aspetto: agli script dei progetti generati tramite l'approccio 3, è stata necessaria una supervisione da parte del candidato. Infatti, sono state eseguite alcune modifiche sugli script generati da chatGPT 4, in cui il candidato ha variato la forma grammaticale, il tono del voiceover rendendolo più intrigante e coinvolgente e meno pubblicitario e adattando gli script a un vasto pubblico di target, che sono i dipendenti dell'azienda Reply, che fruiscono i video all'interno della piattaforma video interna all'azienda.

5.2.2 Storyboard e Image Generation: DALL-E 3 e Midjourney

La fase successiva alla generazione degli script testuali, relativi ai video AI sviluppati e generati per rappresentare i progetti aziendali, è riguardante la generazione e creazione di immagini o frame video, inseriti all'interno dei prodotto audiovisivi.

Nello specifico, sono state generate, in una prima fase, delle immagini utili a rappresentare, in maniera visiva e approssimativa, il video da realizzare. Successivamente, queste sono state posizionate in serie per generare uno storyboard visivo, relativo al video AI aziendale.

La scelta di creare delle immagini di storyborad, si è rivelata essere utile ed essenziale perché ha permesso al candidato di definire e studiare visivamente il video che si desidera generare, sia in termini di successione di immagini o frame video da mostrare e realizzare che in termini di narrazione e sequenza logica delle vicende.

Attraverso questa breve ma importante fase, il candidato è stato facilitato per la successiva fase di generazione di immagini. Infatti, quest'ultima fase ha permesso di ottenere, effettivamente e in maniera concreta, i frame video finali che compongono i video AI da produrre.

La generazione delle immagini degli storyboard, relativi ai video dei progetti aziendali, è avvenuta immediatamente dopo la fase 1 del workflow.

Una volta approvato e scelto l'approccio 3, come il migliore tra i diversi approcci testati, sono stati presi in considerazione gli script ottenuti tramite il suo utilizzo. Si intuisce che, per i progetti aziendali trattati, è stato utile considerare gli script generati e suddivisi in diverse scene video, per creare le immagini degli storyboard dei video AI.

Per generare le immagini degli storyboard, il candidato ha adottato gli applicativi AI chatGPT 4 e DALL-E 3, che rappresenta il tool AI di Image Generation implementato all'interno della versione 4 del tool di OpenAI. Questo tool integrato ha permesso la generazione di immagini, in formato quadrato, di ottima qualità.

La generazione delle immagini, relative agli storyboard, è stata eseguita nel seguente modo:

- Sono stati selezionati in input, i diversi script (distribuiti nelle scene video) ottenuti nello step 1 del workflow, tramite l'approccio 3;
- Dalle frasi indicanti gli script prodotti, inserite in ingresso, è stato chiesto a chatGPT 4 di generare delle immagini in output, che fossero vicine il più possibile alla descrizione testuale degli script, inserita in input.
  Nella generazione effettuata, il candidato ha controllato che le immagini ottenute fossero in linea al contenuto testuale delle frasi indicanti gli script.

Esempio: considerato un certo progetto, è stato copiato il contenuto di uno script generato (e quindi è stata selezionata e copiata la frase che rappresenta lo script) all'interno dell'interfaccia di chatGPT 4 e, in seguito, è stato chiesto di trasformare questo contenuto testuale in un'immagine, adatta per uno storyboard di video. Praticamente, è stata svolta una generazione di immagine di tipo: text-to-image generation.

Qui di seguito è mostrata un'immagine di storyboard, relativa a un video aziendale creato, ottenuta tramite text-to-image generation, in cui si sono unite le funzionalità di chatGPT 4 e DALL-E 3.

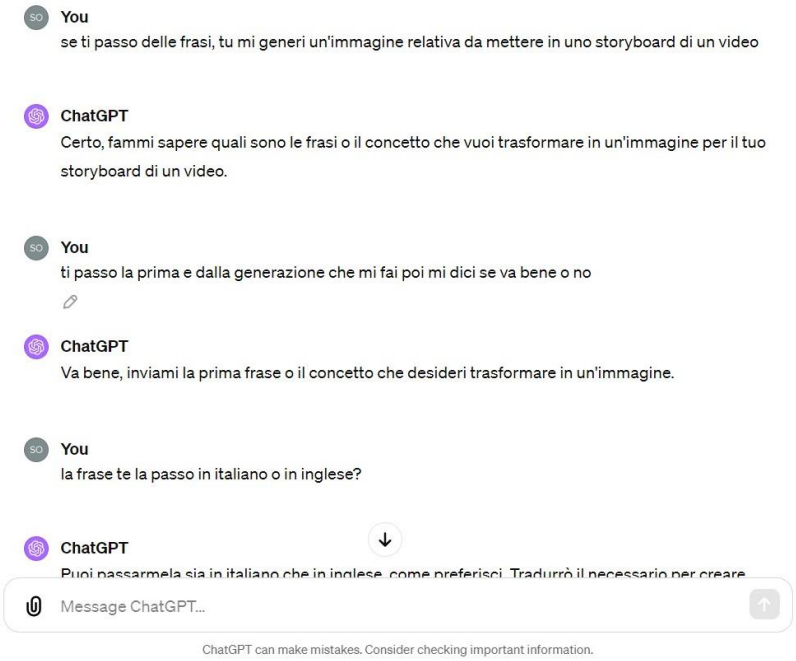

Figura 5.10 Richiesta di generare un'immagine, per lo storyboard del video AI, a partire da descrizione testuale inserita in input

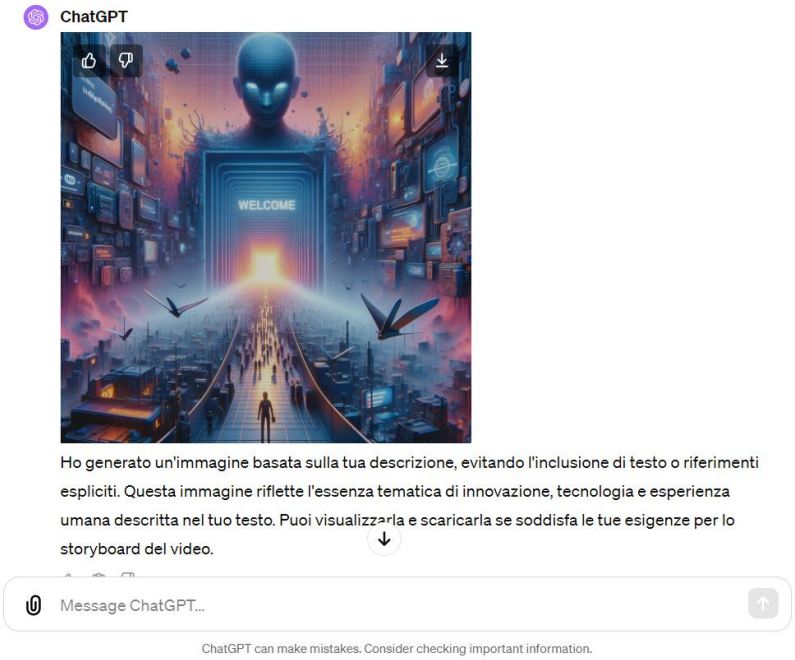

Figura 5.11 Immagine per storyboard video AI, generata tramite DALL-E 3, integrato in chatGPT 4 (sotto è mostrato un commento generato dal tool AI)

Dalla fase di generazione di immagini relative agli storyboard dei video, è stata eseguita la fase di Image Generation. Questa è consistita nella realizzazione delle immagini effettive del video AI da realizzare. Le immagini o i frame video sono stati generati tramite il tool AI di Image Generation, Midjourney.

Il procedimento seguito per la generazione delle immagini dei video è stato articolato ma preciso.

A livello pratico, una volta generate le immagini degli storyboard dei video aziendali, queste sono state prese come punto di riferimento per la creazione dei frame video effettivi dei video. Rispetto alle immagini degli storyboard, in cui ognuna è relativa a una scena video contenente uno specifico script (esempio se il progetto è stato scomposto in 11 scene video, allora le immagini di storyboard create sono 11, una per ogni scena video che contiene uno script testuale), le immagini effettive dei video sono molte di più. Questo è dovuto al fatto che, ogni script generato, per ciascuna scena video, contiene un certo numero di righe e, a queste, corrisponde una determinata immagine generata. Da questa osservazione, si comprende come il numero di immagini, prodotte per una relativa scena video e, dunque, per un intero script testuale narrante, varia dalle 2 alle 4 generazioni.

Per giungere alla generazione finale delle immagini è stato seguito questo procedimento:

- E' stata inserita in input, all'interno di Midjourney, l'immagine di storyboard relativa a una scena video.
- L'immagine di storyboard è stata inserita mediante il comando "/describe" fornito da Midjourney. Questo comando ha permesso al candidato di generare 4 prompt testuali (o 4 descrizioni testuali) differenti, che rappresentassero l'immagine data in ingresso. In breve, è stata svolta un'operazione, a ritroso, di image-to-text generation, in cui il termine "text" inteso indica i prompt testuali generati, che meglio descrivono l'immagine data in ingresso.
- Una volta ottenute le quattro versioni diverse di prompt testuali, che indicano il contenuto dell'immagine passata in input, sono state generate le immagini in output, a partire dai prompt testuali di input creati da Midjourney. Per fare queste generazioni di tipo text-to-image, il candidato ha premuto sui bottoni presenti, nella UI di Midjourney, subito dopo l'immagine di storyboard caricata.

- Dalle generazioni di immagini (quadrate) ottenute dai 4 prompt testuali, il candidato ha scelto le versioni migliori, che meglio rappresentassero lo stile e il significato da adottare nel video da generare.
- Scelte le migliori generazioni di immagini, è stato necessario selezionare e copiare il prompt testuale, relativo alle 4 immagini quadrate, e riadattarlo alle proprie esigenze stilistiche (3D in stile motore grafico Unreal Engine) e di formato (16:9).
- Dal prompt testuale adottato e incollato, si sono ottenute generazioni di 4 immagini alla volta, che appaiono nello stile e nel formato desiderati dal candidato.
- Da queste generazioni di immagini, e dopo un numero di tentativi, sono state raccolte una serie di immagini che meglio rappresentano la scena video. Ovviamente, ritornando alla questione del numero di immagini da generare per ogni scena video, sono state svolte tante generazioni di immagini per poter scegliere più immagini valide che indicano in maniera esaustiva la scena video. Quindi, in base a quelle che sono le diverse frasi che compongono un certo script completo e, dunque, una certa scena video, sono state prodotte delle immagini associate a ciascuna frase.

Spiegato il procedimento generale applicato, questo è stato replicato per ciascuna immagine di storyboard, generata nella fase precedente all'Image Generation effettiva (ovvero nella fase di generazione di immagini approssimative per gli storyboard dei video).

È utile considerare che il candidato, per qualche generazione di immagine da adottare nei video AI, ha deciso di adottare la classica tipologia di generazione: text-to-image generation, per ottenere immagini intermedie tra una scena video e l'altra, ma anche per testare l'applicativo AI Midjourney in termini di affidabilità e precisione nella realizzazione di contenuti visivi, come le immagini.

Qui di seguito sono mostrate alcune schermate che spiegano, visivamente, il procedimento applicato:

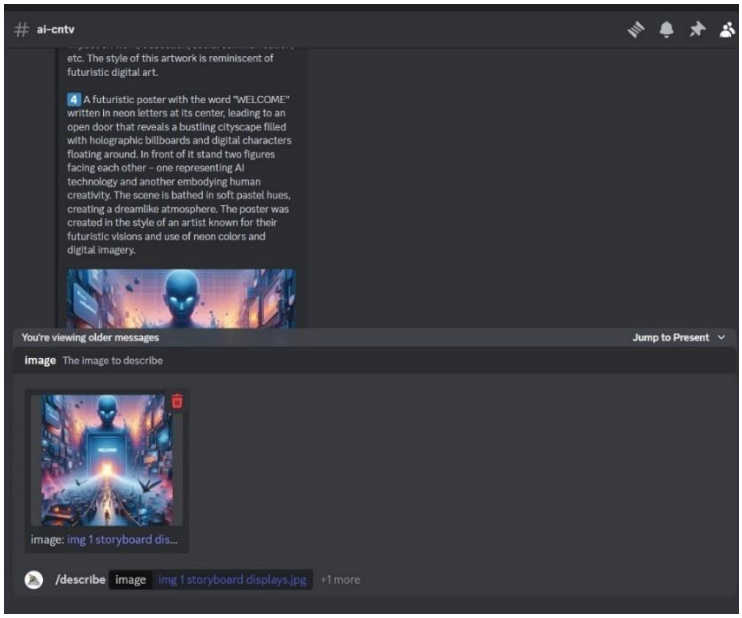

Figura 5.12 Applicazione del comando "/describe" con l'inserimento in input di un'immagine di storyboard, generata e relativa a un progetto aziendale scelto

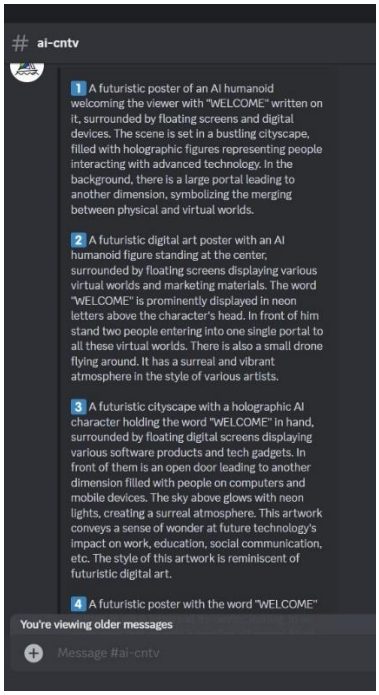

Figura 5.13 Versioni di script testuali generati a partire dall'immagine di storyboard inserita in input (il quarto prompt è mostrato nella schermata precedente)

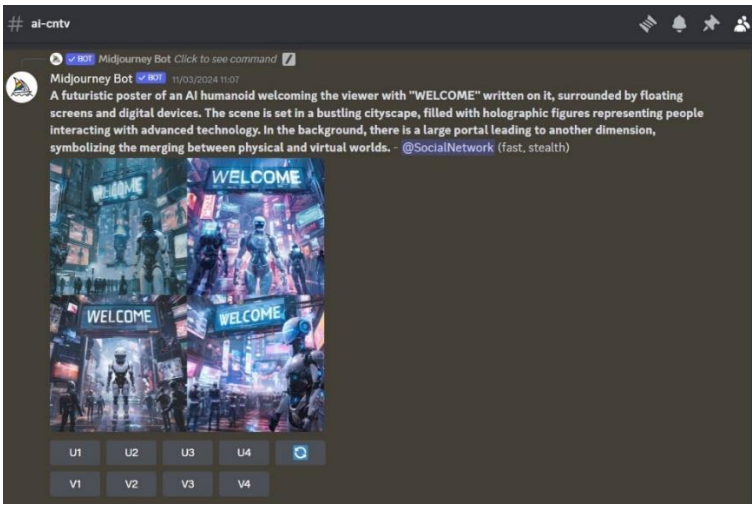

Figura 5.14 Text-to-image generation (di immagini quadrate) ottenuta a partire da una versione di prompt testuale prodotta (si tratta della versione 1 del prompt testuale generato da Midjourney, visibile nella schermata 5.13)

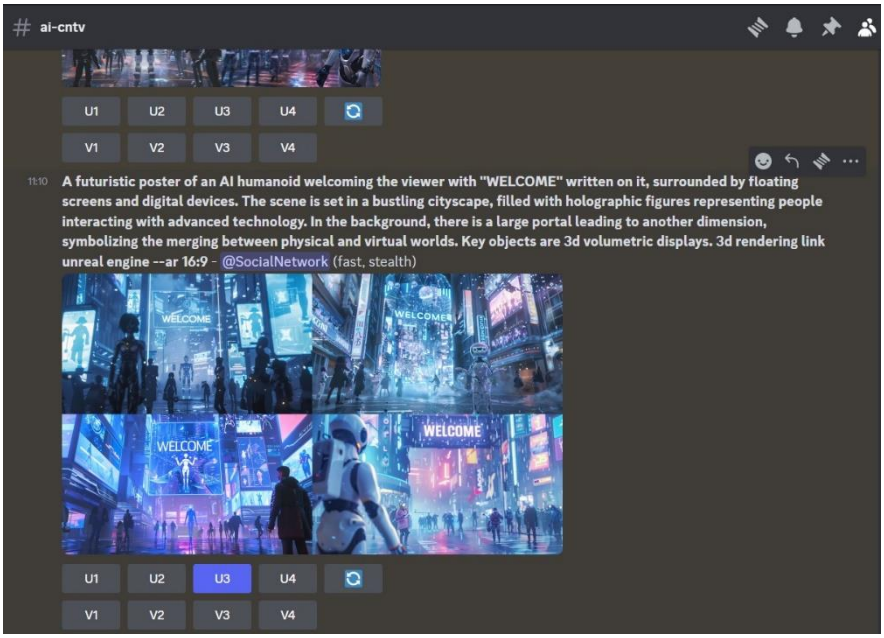

Figura 5.15 Generazione di immagini ottenuta dopo aver copiato e incollato il prompt testuale mostrato nella schermata precedente, su cui sono stati inseriti altri dettagli per soddisfare le esigenze stilistiche e di formato del candidato (3d rendering link unreal engine e –ar 16:9)

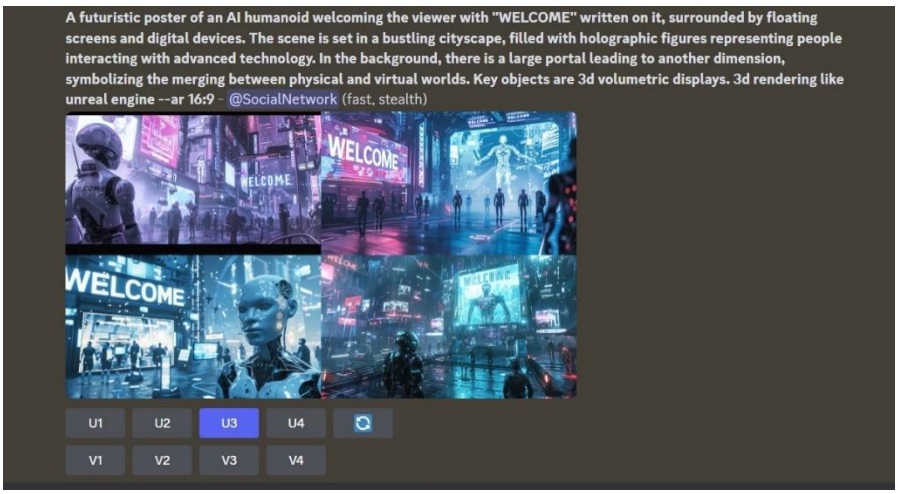

Figura 5.16 Altra generazione di immagini ottenuta dallo stesso prompt testuale raffigurato nella schermata precedente, a meno di un termine ("link unreal engine" nella schermata precedente e "like unreal engine" in questa schermata)

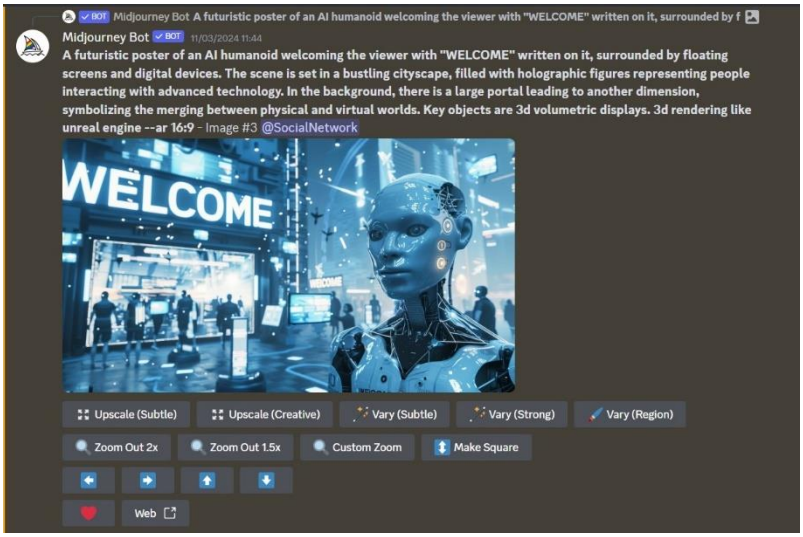

Figura 5.17 Immagine ottenuta dall'applicazione dell'opzione di Upscaling di Midjourney, nella quale si è ottenuta un'estensione e un miglioramento della risoluzione dell'immagine selezionata nella schermata precedente (indicata dal pulsante U3 che è stato cliccato per generare questa versione ingrandita di immagine). L'immagine mostrata rappresenta uno dei frame video inseriti in un video AI generato

Inoltre, durante la generazione di alcune immagini utilizzate nei video AI realizzati, sono state applicate delle operazioni di "Vary Region" utili a trasformare una determinata area di immagine. Il candidato ha applicato l'operazione di variazione di regione per apportare modifiche in alcune porzioni di frame ricreato.

Nelle schermate che seguono si può notare l'applicazione del "Vary Region".

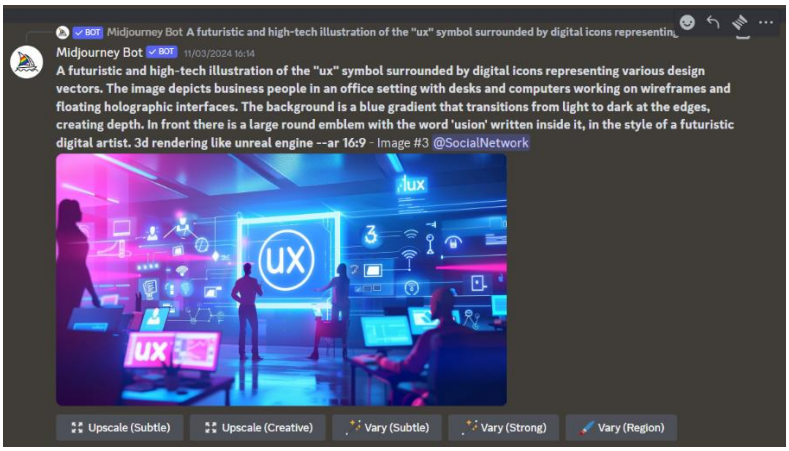

Figura 5.18 Generazione di immagine (ottenuta dal prompt testuale) sulla quale è stata applicata l'opzione di Vary Region (osservabile alla destra schermata)

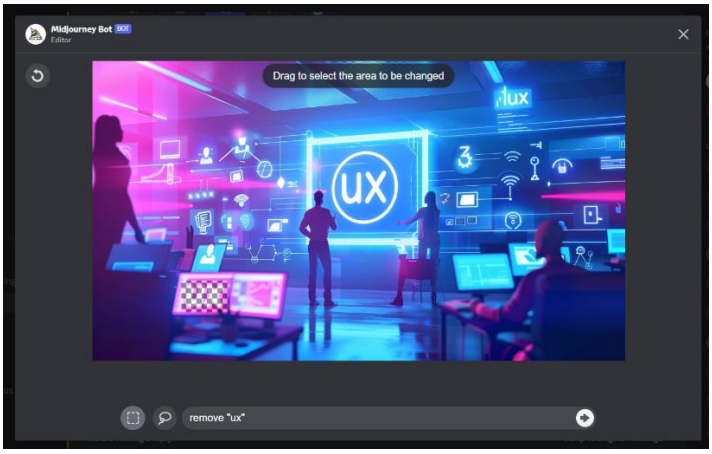

Figura 5.19 Interfaccia disponibile dopo il click sul button "Vary Region". Nell'immagine è stata selezionata l'area dello schermo (visibile con la griglia in scala di grigi in basso a sinistra) su cui rimuovere la scritta "ux", come indicato nel campo di scrittura.

Figura 5.20 Immagine ottenuta dopo l'applicazione della variazione di regione, con lo schermo privo della scritta "ux"

In questo caso, è stato necessario effettuare un'ulteriore operazione di Vary Region relativa alla rimozione della sagoma femminile presente alla sinistra dell'immagine. Quindi, a una prima applicazione dell'opzione di variazione area immagine, ne è stata eseguita una seconda i cui risultati sono visibili nelle prossime due catture schermo mostrate.

Figura 5.21 UI di Vary Region con area selezionata e indicante la porzione di immagine da rimuovere, relativa alla sagoma femminile. Espressione comando remove "women" nell'area di testo.

Figura 5.22 Immagine finale generata e ottenuta senza gli elementi trattati nel Vary Region ("ux" sullo schermo e sagoma femminile)

5.2.3 Video Generation: Runway

La terza fase del workflow, sviluppata dal candidato, è quella relativa alla Video Generation. Questa è una delle fasi principali, se non la più importante, del flusso di lavoro rispettato per svolgere l'altra parte del progetto di tesi (ovvero la realizzazione dei video AI legati ai progetti aziendali scelti).

Per eseguire la Video Generation, è stato utilizzato il tool AI Video Generativo, RunwayML.

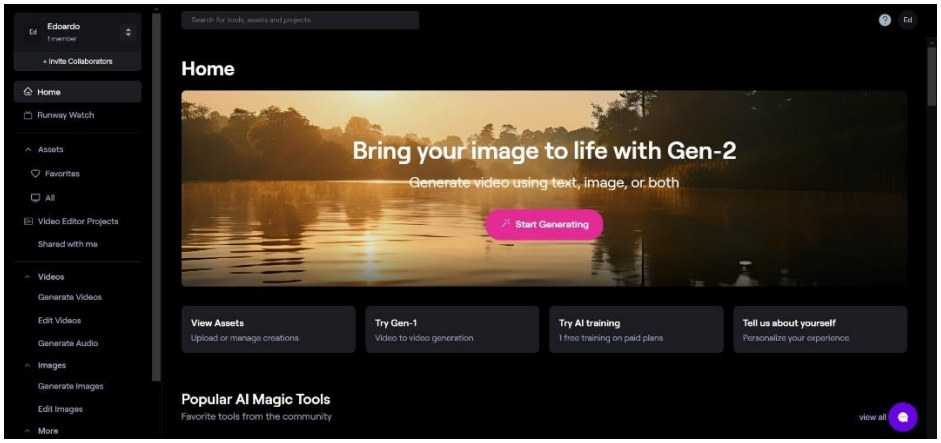

Figura 5.23 Homepage del tool AI di Video Generation: RunwayML

Una volta ottenute le diverse immagini, generate nello step 2 di Image Generation del workflow, è stata sviluppata la fase di generazione dei video AI. In particolare, sono state acquisite le immagini (i frame video) prodotte e sono state create le diverse clip o scene video. Per realizzare le diverse scene video AI, sono state eseguite delle animazioni sulle immagini ottenute con Midjourney.

Principalmente, l'animazione, e la relativa clip video generata per ciascuna immagine, è stata ottenuta mediante l'operazione di image-to-video generation. Si intuisce come, le diverse scene video AI, sono state ottenute in output a partire da immagini inserite in input.

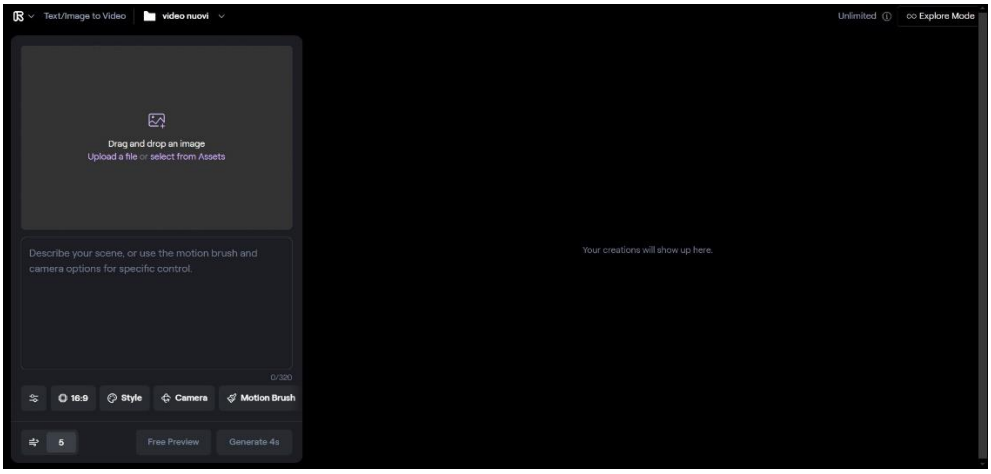

Figura 5.24 Schermata che mostra l'interfaccia di RunwayML su cui all'interno si creano le generazioni di video AI. In alto sinistra, si inserisce l'immagine da generare in video; in basso a sinistra, si inserisce il testo da trasformare in video. Riempiendo entrambi i campi, si può ottenere una text+image-to-video generation

All'interno della sezione di Runway mostrata sopra, il candidato ha prodotto le diverse generazioni dei video AI. Infatti, le diverse immagini prodotte in Midjourney, sono state inserite nell'apposito campo, mostrato in alto a sinistra nella schermata caricata.

Una volta eseguito l'upload dell'immagine, sono state considerate e applicate una serie di operazioni, utili a sviluppare le generazioni AI delle clip video, cambiate in base a una determinata immagine inserita in input.

Ogni operazione di generazione, è stata svolta mediante l'utilizzo di uno o più bottoni visibili nello screenshot 5.19 (in basso a sinistra), e presenti nella UI di Runway, che consentono di eseguire diverse operazioni e offrono diverse funzionalità video generative.

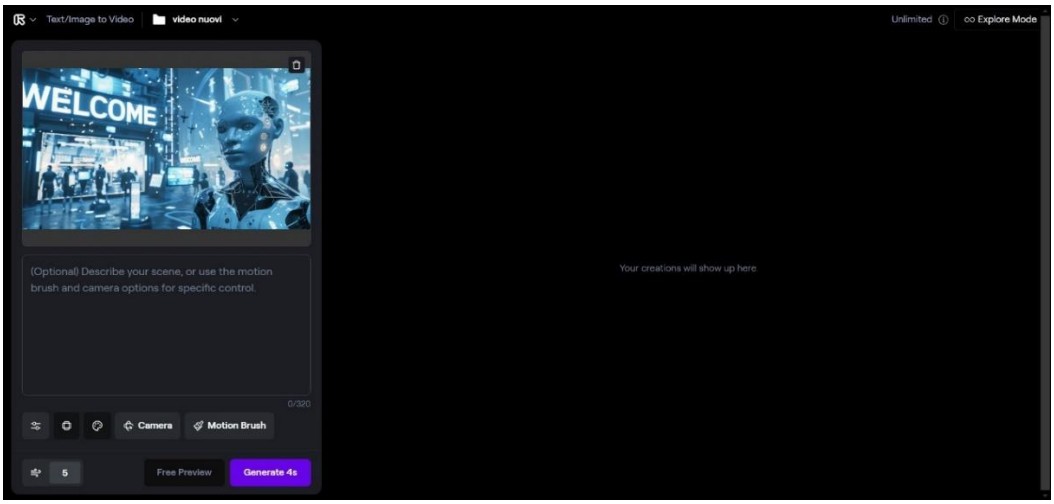

Figura 5.25 Inserimento in input di un'immagine, generata tramite Midjourney, che è pronta ad essere trasformata in video

Dall'inserimento di un'immagine, si nota come viene evidenziato il pulsante "*Generate 4s*", che indica l'operazione di generazione dell'animazione dell'immagine, per ottenere, dunque, una clip video composta dall'immagine animata. Dunque, è il pulsante che viene premuto per produrre ogni generazione video AI.

Il pulsante citato, se premuto senza eseguire altre operazioni, genera un'animazione automatica, sia nei movimenti di camera adottati che nei movimenti da attribuire ai diversi oggetti e soggetti che

compaiono nell'immagine. Dalla denominazione del bottone, si capisce come le singole clip video AI generabili presentano una durata iniziale di 4 secondi.

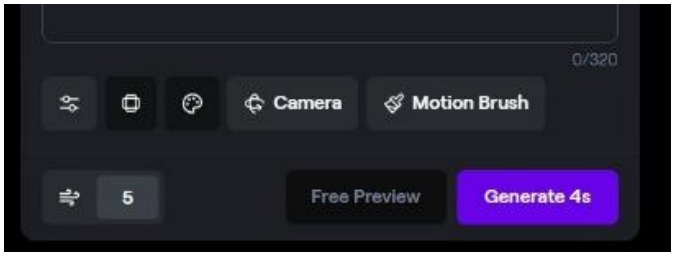

Figura 5.26 I bottoni della schermata precedente mostrati in maniera più evidente, tra cui compare "Generate 4s"

Oltre al comando di Generate dei video AI, emergono i bottoni, e i relativi comandi, che sono stati utili al candidato, per gestire e generare i movimenti e le animazioni delle immagini, che hanno permesso di ottenere le generazioni finali delle diverse clip video AI, relative ai progetti aziendali.

Il candidato ha spiegato, nelle prossime righe, le diverse funzionalità video generative che ha adottato nella realizzazione delle clip o scene video AI.

Il bottone "*Camera*" consiste nella scelta dei diversi movimenti di camera che, eventualmente, si vogliono applicare, per produrre specifiche animazioni delle immagini inserite in input.

Al click del pulsante, compare una schermata denominata "Camera Motion" (tradotto movimento di camera) che offre 6 tipologie differenti di movimenti da assegnare alla camera, durante la ripresa dell'immagine, che verrà animata seguendo il camera motion selezionato.

Oltre a specificare il movimento, è possibile gestire l'intensità con cui la camera può svolgere un determinato movimento. L'intensità, in altre parole, indica quanto si vuole accentuare e mettere in evidenza un determinato movimento di camera: se generarlo in maniera piuttosto evidente oppure se mantenerlo lieve.

Il candidato, nella schermata mostrata sotto, una volta inserita un'immagine in ingresso, presente e animata per un video AI realizzato, ha selezionato il movimento di camera "Pan", che applica alla camera un movimento di inclinazione lungo l'asse x/orizzontale (opposto al movimento "Tilt").

Nello specifico, il movimento di camera "Pan" è stato applicato con un valore di intensità pari a 2, muovendosi verso l'asse positivo dell'interfaccia (quindi spostando la leva verso destra). Se invece si decide di spostare la leva verso sinistra, allora il valore del movimento di camera diventerà più piccolo, fino ad essere negativo, e il movimento della camera tenderà a comportarsi in maniera opposta rispetto al movimento selezionato dal candidato.

Inoltre, è possibile digitare direttamente il valore di intensità che si vuole assegnare al movimento di camera, inserendolo nel campo mostrato, in questa schermata, dal contenuto "2.0". L'interfaccia presenta questo valore a causa del movimento di camera applicato dal candidato sull'immagine inserita in input, ma il camera motion è facilmente modificabile, cliccando nelle apposite aree e inserendo il valore numerico che si desidera associare al movimento.

Una terza possibilità di assegnazione del valore di movimento di camera, è quello relativo al click dei pulsanti che mostrano l'icona del movimento da applicare. Nella schermata presente sotto, una di queste due icone è riempita con un colore viola e questo indica che il candidato ha effettuato un movimento di camera associato alla raffigurazione dell'icona presente a destra, che raffigura il movimento "Pan" effettuato verso una certa inclinazione.

Se si decide di cliccare l'altra icona, il valore del movimento di camera viene decrementato fino a diventare, eventualmente, negativo e la leva tende a spostarsi verso sinistra. Al click sull'icona di sinistra, questa diventerà viola mentre l'altra non avrà più il colore di riempimento.

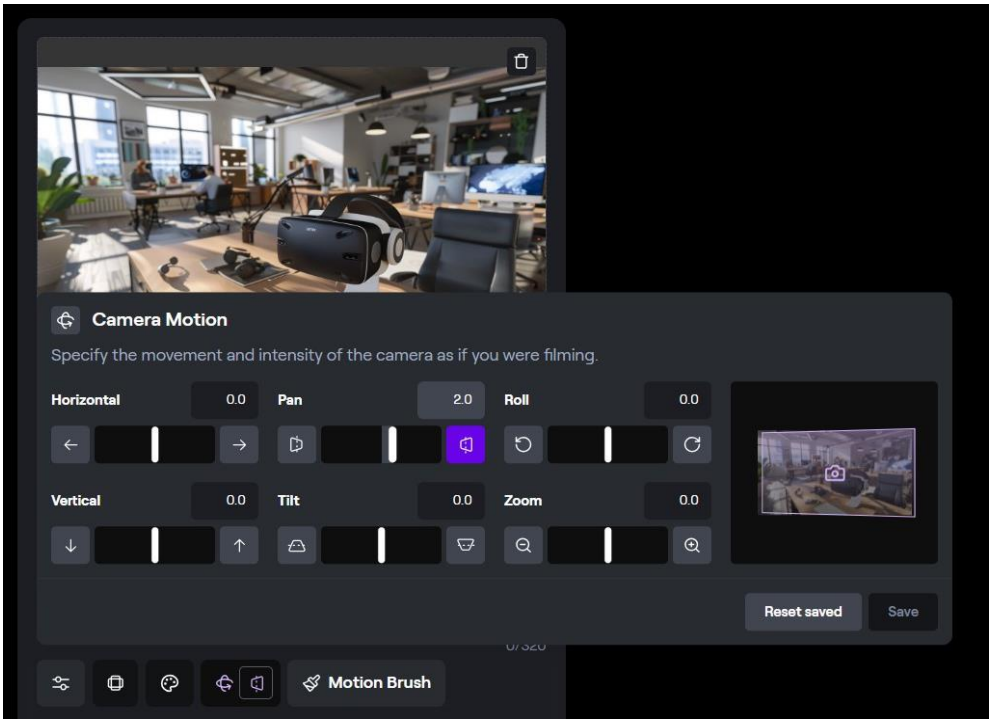

Figura 5.27 Interfaccia di gestione dei movimenti di camera, da assegnare all'animazione dell'immagine. Quest'ultima è stata inserita in input ed è stata utilizzata in un video AI realizzato. Sono visibili le tipologie dei movimenti di camera, tra cui quello adottato dal candidato per generare l'animazione

Un altro elemento, utile al candidato per applicare correttamente il movimento di camera, è quello presente alla destra della schermata, in cui il tool consente di far visualizzare l'anteprima del movimento di camera, applicato sull'immagine in miniatura.

In particolare, nel caso della schermata, si nota come il movimento "Pan" con valore 2.0 viene applicato sull'immagine attraverso una leggera inclinazione, come mostrato dall'area in viola sovrapposta all'immagine. Se si desidera resettare le modifiche apportate, è possibile cliccare il bottone "Reset saved" che consente di resettare tutte le operazioni di movimento camera effettuate.

Oltre al movimento di camera "Pan", sono applicabili altre 5 tipologie di movimenti tra cui:

- "Horizontal" che consiste nel traslare la camera virtuale di Runway lungo l'asse x, in cui è possibile decidere se effettuare il movimento di camera orizzontalmente verso sinistra o verso destra, a seconda dello spostamento della leva o in base al valore numerico inserito nel campo o, per ultimo, in base al click del bottone raffigurante l'icona del movimento da effettuare;
- "Vertical" che consiste di traslare la camera virtuale di Runway lungo l'asse y, in cui il la camera svolgerà movimenti verticali verso l'alto o verso il basso, a seconda dei valori che si decidono di assegnare al camera motion selezionato;
- "Tilt" che permette di inclinare la camera lungo l'asse y/verticale e, i relativi effetti di movimento dipendono dai valori inseriti negli appositi campi dell'interfaccia;
- "Rool" che permette una rotazione della camera in senso orario o antiorario, a seconda del valore di movimento che si decide di assegnare al tool;
- "Zoom" che consente alla camera di avvicinarsi all'immagine selezionata, immergendosi dentro l'ambiente costituito dall'immagine oppure di allontanarsi gradualmente dall'immagine, e dall'ambiente da essa mostrata.

Altro bottone rilevante dell'interfaccia, e utile per le generazioni di video AI effettuate, è quello di "*Motion Brush*", posizionato di fianco al pulsante "*Camera*".

Questa funzionalità consente di regolare i movimenti di determinati oggetti e soggetti che compaiono all'interno dell'immagine inserita, assegnando a questi, specifici movimenti e determinate animazioni, lungo certe direzioni. L'operazione di Motion Brush è utile perché ha permesso al candidato di animare differenti elementi dell'immagine, enfatizzandoli all'interno della clip video AI generata e, mostrandoli in maniera evidente, all'interno dell'intero video AI realizzato.

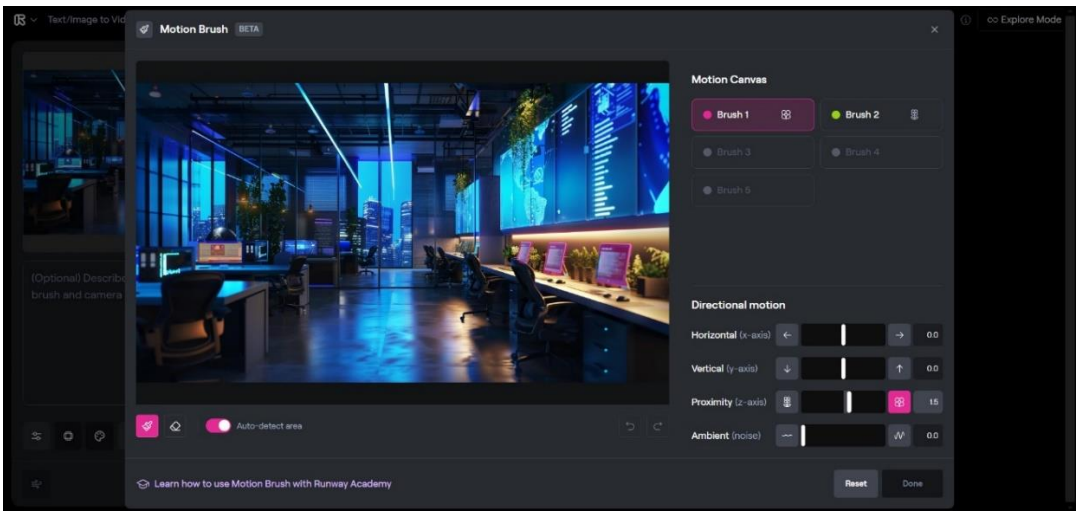

Figura 5.28 Applicazione del Motion Brush sull'immagine selezionata in input e utilizzata in un video AI realizzato

Nella schermata si nota un'applicazione della funzionalità di Motion Brush a cui, al click dell'apposito pulsante presente nella UI generale, compare un'interfaccia dedicata. Nel dettaglio, si nota l'immagine inserita in ingresso (usata in un video AI prodotto), che si desidera animare, e sono rilevanti alcuni elementi che sono stati evidenziati dal candidato, attraverso l'applicazione del motion brush sulle superfici di questi.

Le parti, sulla quale è stato applicato l'effetto di motion brush, sono rappresentate dagli schermi dei computer.

Quelli di destra sono soggetti a un'applicazione del motion brush che avviene in un movimento direzionale (o di direzione) rivolto in prossimità dell'asse z (profondità), in cui in particolare si decide di far emergere leggermente il contenuto dagli schermi. L'applicazione del motion brush su questi elementi dell'immagine è rappresentata dal colore viola.

Mentre, lo schermo del computer posto a sinistra dell'immagine, presenta un'applicazione del motion brush indicata dal colore verde e, come si può notare dalla sezione Motion Canvas della schermata (con la presenza della voce "Brush 2" affiancata da un cerchio verde e da un'icona di un fiore che porta all'idea di allontanamento), si intuisce che il contenuto mostrato nel display è stato animato sempre in prossimità dell'asse z, ma con un effetto opposto a quello applicato sugli altri schermi dei computer presenti a destra dell'immagine, indicati dal colore viola.

Infatti, nella schermata, la sezione Motion Canvas mostra le diverse applicazioni di motion brush applicate dal candidato, in cui da un lato si notano le applicazioni viola del motion brush denominato "Brush 1", affiancato da cerchio viola e icona del fiore ingrandita, e dall'altro lato emerge l'altra applicazione di motion brush denominata "Brush 2", affiancata da cerchio verde e icona del fiore intero.

L’area “Directional motion” è quella in cui è possibile assegnare il movimento direzionale lungo cui il motion brush deve avere affetto. Quindi, selezionati certi oggetti attraverso il brush di movimento, è utile specificare in quale direzione deve avere effetto l’animazione e, dunque, il movimento degli elementi selezionati. Si hanno 4 opzioni di movimenti direzionali gestibili con il motion brush:

- “Horizontal” consiste nello spostamento, nel movimento e nell’animazione dell’area di oggetto selezionata, lungo l’asse x (verso destra o verso sinistra);
- “Vertical” possiede una logica opposta a “Horizontal”;
- “Proximity” permette all’area selezionata di emergere o penetrare rispetto alla superficie, dando l’idea di prossimità lungo l’asse z, rivolta nei due punti di vista (ingrandimento o allontanamento);
- “Ambient” permette di inserire un movimento direzionale nell’area di immagine selezionata, caratterizzato da rumore (noise) di animazione, applicabile verso direzioni che mostrano un rumore poco accentuato (leva verso sinistra) o molto accentuato (leva verso destra). Con questa opzione è come se di decidesse di distorcere, per l’area di immagine selezionata, il contenuto in maniera poco o tanto rilevante.

Nella parte in basso a sinistra della schermata, sono presenti i bottoni di “pennello” del motion brush, che permette di selezionare le aree su cui applicare l’effetto di movimento e animazione (nel caso della schermata inserita e riempito di viola ed è stato, dunque, utilizzato) e l’icona di “gomma” che consente di cancellare, eventualmente, le aree colorate su cui è stato applicata la funzionalità di motion brush.

La selezione e deselezione delle aree soggette al motion brush, possono essere gestite in maniera automatica tramite l’opzione di “Auto-detect area” (attivata dal candidato sull’immagine inserita in input e mostrata attiva all’interno della schermata con lo slider rivolto verso destra) su cui, al click sulla superficie dell’oggetto, il riempimento di colore e l’individuazione dell’area avviene automaticamente.

Nel caso dell’immagine inserita, è stato utile utilizzare e attivare l’opzione di “Auto-detect area”, in quanto gli schermi dei pc sono abbastanza omogenei come forme, e per facilitare il lavoro di riempimento dell’area su cui applicare il motion brush, è stato opportuno decidere per questa modalità.

Se si decidesse di non attivare l’opzione di “Auto-detect area”, la selezione delle superfici degli oggetti deve essere eseguita a mano e in maniera precisa. Per esempio, il candidato ha applicato questa decisione, in alcune immagini inserite in input e presenti nei video AI realizzati in cui, alla presenza di una mano composta da dita, è stato necessario selezionare in maniera fine le dita, per animarle e farle muovere in maniera il più possibile efficiente. Se si seleziona la mano con l’auto detect c’è il rischio di selezionare anche altre parti di immagine, vicine alle dita o al palmo, e produrre animazioni e movimenti non desiderati.

Inoltre, se si usa la selezione delle aree manuale e non automatica, compare, nella UI, la possibilità di scegliere la grandezza del pennello e della gomma per eseguire, rispettivamente, delle selezioni e deselezioni di aree in maniera più grossolana o precisa. Infatti, è possibile selezionare il pennello grande o piccolo per colorare le aree in modo superficiale o accurato oppure selezionare la gomma grande o piccola per cancellare le aree, inizialmente selezionate e colorate, in modo superficiale o accurato.

Altre applicazioni di Motion Brush, sono state applicate su diverse immagini passate in input, per regolare e gestire l’animazione desiderata.

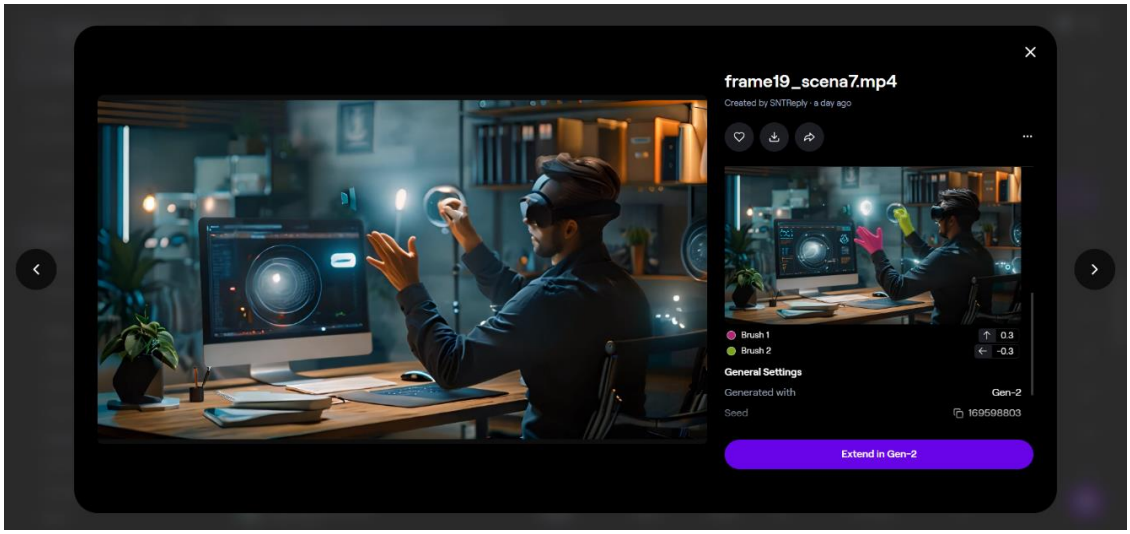

Figura 5.29 Avvenuta applicazione del motion brush sull'immagine inserita e utilizzata in un video AI

Nella presente schermata si notano, nella parte destra, le aree selezionate dell'immagine, con i diversi motion brush (viola e verde) che indicano differenti animazioni e movimenti di direzione applicati sull'immagine, che è stata animata. L'animazione è mostrata nella parte sinistra della schermata, su cui è stato catturato un frame di movimento.

Infatti, la mano destra della figura, colorata di verde, è stata animata e mossa orizzontalmente verso sinistra con un valore pari a -0.3, mentre la mano sinistra, colorata di viola, è stata animata e mossa verticalmente verso l'alto con un valore pari a 0.3.

Da sottolineare, l'opzione di estensione di ulteriori 4 secondi, della scena o clip video ottenuta. Questa estensione è possibile attraverso il click sul pulsante "Extend in Gen-2", in cui l'applicativo Runway consente ulteriori modifiche alla scena video creata, prolungandola nella durata.

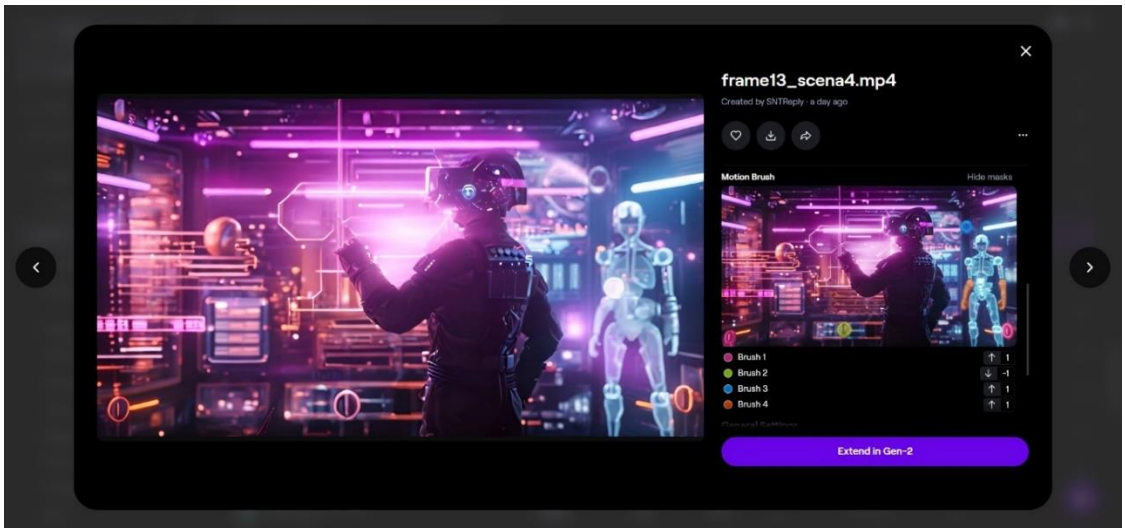

Figura 5.30 Altra avvenuta applicazione del motion brush su un'altra immagine inserita e utilizzata in un video AI

Anche in questa schermata, è mostrata l'applicazione del motion brush da parte del candidato. Nello specifico, sono state adottate 4 tecniche differenti di motion brush, come si può osservare dalla parte destra della schermata. Infatti, le diverse aree selezionate e colorate in diverso modo all'interno dell'immagine inserita in input, rappresentano le zone in cui sono state applicate specifiche animazioni con diversi movimenti direzionali eseguiti.

I 4 tipi di motion brush sono indicati da:

- Un motion brush, denominato “Brush 1” e indicato dal viola, che esegue un’animazione verticale degli elementi, che si muovono verso l’alto, con un valore numerico pari a 1;
- Un motion brush, denominato “Brush 2” e indicato dal verde, che esegue un’animazione lungo l’asse y degli elementi, che si spostano verso il basso, con un valore numerico pari a -1;
- Un motion brush, denominato “Brush 3” e indicato dal blu, che esegue un’animazione verticale degli elementi, che si muovono verso l’alto con valore numerico posto a 1;
- Un motion brush, denominato “Brush 4” e indicato dall’arancione, che consente animazioni e movimenti uguali a quelli del “Brush 3”.

Le applicazioni dei motion brush utilizzati, sono visibili nella parte sinistra della schermata, in cui è stato catturato uno specifico frame video di animazione.

Anche in questa cattura immagine è visibile il pulsante “Extend in Gen-2”, per consentire l’estensione e la modifica della scena video generata.

Altri bottoni rilevanti nell’interfaccia utente, visibile nella figura 5.21, sono quelli relativi a “*General motion*” e a “*Style*”.

Il pulsante “General motion”, visibile con il valore numerico inserito e pari a 5, consiste nell’animare l’immagine con una velocità o intensità di movimento pari a un valore indicato nel campo e scelto dall’utente. Nella schermata è mostrato il valore 5 (valore intermedio tra 0 e 10, che sono i valori massimi e minimi consentiti da Runway) che è quello di default di Runway, quando si accede alla schermata di video generazione AI.

Il candidato ha applicato l’uso di questa funzionalità, per alcune delle scene video AI visibili nei prodotti audiovisivi finali, assegnando valori pari a 5, 6 e 7.

Quindi, più questi sono alti, maggiore movimento sarà generato nell’immagine inserita in input, mentre più i valori sono piccoli, inferiore risulterà essere il movimento e l’animazione generata nell’immagine di input.

L’utilizzo della funzionalità di General motion dipende ed è vincolata dall’utilizzo delle funzioni di Camera Motion e Motion Brush. Infatti, se si utilizzano uno o entrambi i campi citati, il valore di General Motion non viene considerato e non può essere adoperato nella generazione, in quanto l’animazione dell’immagine dipende strettamente dalla gestione dei movimenti di camera e, nel caso venga utilizzato, dal motion brush che sviluppa animazioni più precise e decise dall’utente.

Se invece, si presentano dei dubbi su quali animazioni generare all’interno dell’immagine inserita in input, allora è utile applicare l’operazione di General Motion, senza utilizzare le altre due funzionalità citate (che vengono disabilitate).

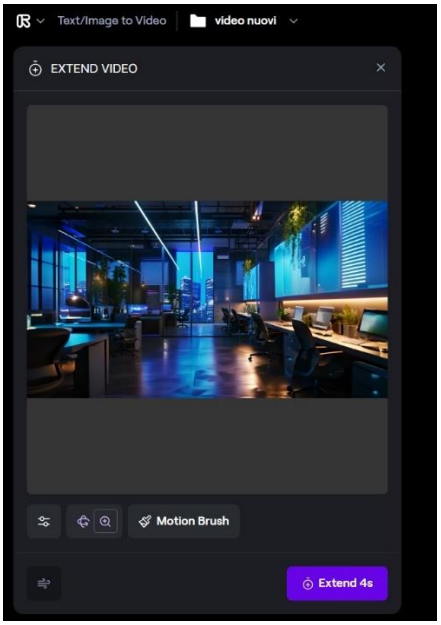

Figura 5.31 Dimostrazione di come il campo General motion scompare, se si utilizzano, nell'immagine inserita in input, le operazioni di motion camera e/o di motion brush

L'esempio pratico di dipendenza tra le 3 operazioni citate, è mostrata nella schermata inserita sopra.

Si nota, come il candidato abbia preferito creare le animazioni nell'immagine inserita, attraverso le operazioni di Camera Motion e Motion Brush, piuttosto che utilizzare la funzionalità di General Motion. Quest'ultima infatti viene disabilitata dall'applicativo AI, in quanto comprende che sono state già utilizzate le altre due operazioni principali di generazione dell'animazione e del movimento.

Questa è la dimostrazione che l'utilizzo delle due operazioni, impedisce l'applicazione dell'altra e viceversa.

L'altro pulsante, presente nella UI della figura 5.21, è quello di "Style", indicato da una tavolozza artistica.

Nelle clip o scene video AI realizzate dal candidato per narrare i progetti aziendali, non è stato utilizzato il campo Style. La motivazione è relativa al fatto che, avendo svolto generazioni di tipo image-to-video, le generazioni dei diversi video AI sono dipese dallo stile già presente all'interno delle diverse immagini caricate in input.

Nel caso dei video AI dei progetti aziendali, lo stile mostrato è il 3D rendering, presente e generato all'interno delle immagini prodotte nella fase di Image Generation dello step 2 del workflow. Quindi, i video AI hanno mantenuto lo stesso stile delle immagini inserite e animate.

Oltre allo stile, anche il formato delle scene video AI ottenute, è dipeso da quello presente e generato nelle immagini create tramite Midjourney. Il formato, delle immagini e dei video AI ottenuti, è il 16:9.

Svolte e spiegate le diverse funzionalità video generative (trattate sia dal punto di vista del progetto di tesi che dal punto di vista generale) adottate dal candidato, per la realizzazione dei video AI tramite Runway, e spiegato il tool nelle sue aree e sezioni, è stato possibile ottenere differenti clip video AI relative ai progetti aziendali.

Una volta prodotte, queste clip sono state inserite e unite all'interno del software di video editing Adobe Premiere, per creare i video AI finali, composti dalle diverse scene video generate in Runway.

5.2.4 Voice e Audio Generation: ElevenLabs

La fase del workflow, successiva all'AI Video Generation, è quella relativa alla Voice e Audio Generation. Questa fase è stata applicata dal candidato, per generare e sintetizzare le voci narranti (o voice over) da adottare nei video AI realizzati. Le voci, inserite nei relativi video AI dei progetti aziendali, sono state aggiunte, insieme alle clip video AI ottenute nello step precedente, all'interno del software di video editing Adobe Premiere.

Ad ogni scena video AI riprodotta, e a ciascuno script elaborato nella fase 1 del workflow, corrisponde una generazione vocale relativa.

Le voci sono state generate tramite il tool AI di Voice e Audio Generation ElevenLabs.

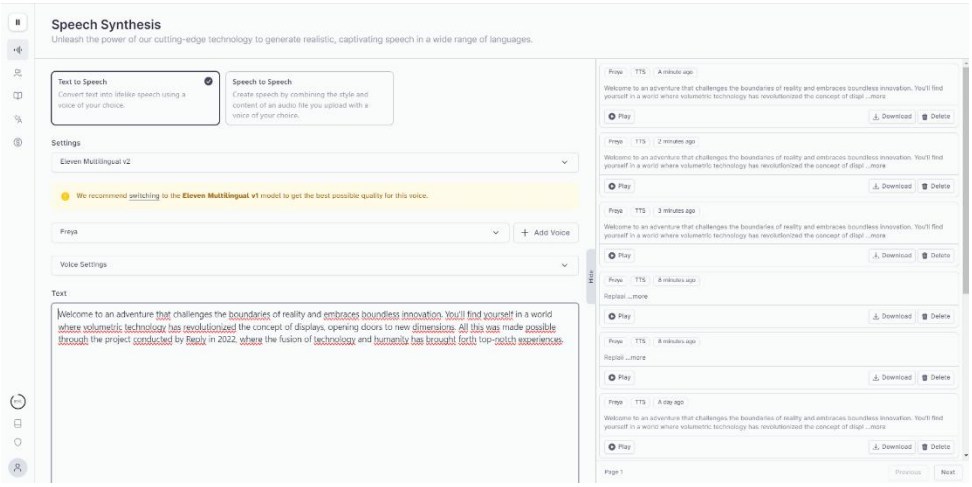

Figura 5.32 Homepage di ElevenLabs con le diverse sezioni. Si nota come la tipologia di generazione selezionata è quella text-to speech (o voice)

La generazione delle voci è stata di tipo text-to-voice generation, in cui è stato inserito in input l'intero script testuale, generato e relativo al progetto aziendale, ed è stata generata in output la relativa voce narrante che replica, in maniera vocale e sonora, il testo scritto negli script, smistati per scene video e sezioni.

Questa quarta fase del worklfow è, dunque, connessa direttamente alla fase 1 del flusso di lavoro.

Infatti, una volta generati (tramite chatGPT4) e uniti i diversi script testuali dei progetti aziendali, questi sono stati passati in input a ElevenLabs per genere la rispettiva generazione vocale.

Il candidato ha inserito gli script testuali uniti tra di loro, all'interno del tool di voice generation, ma per alcune generazioni vocali è stato utile inserire le singole frasi degli script in maniera individuale, per ottimizzare le generazioni vocali.

Le generazioni vocali, effettuate per la realizzazione dei video AI dei progetti aziendali, sono state svolte scegliendo due voci campioni dalla libreria offerta dall'applicativo.

In particolare, le due voci scelte sono una femminile e una maschile, in modo tale da variare nella narrazione vocale. Per le due voci scelte, come anche per le altre voci in generale, si notano i dettagli relativi che identificano la voce. Infatti, nel caso della voce denominata "Freya" si nota come questa sia in lingua inglese americana, presenta uno stile "overhyped" e presenta un tono per narrazione di videogiochi (dunque per i video AI relativi ai progetti aziendali è adatta in quanto il tono testato è in

linea con la narrazione che si vuole dare). L'altra voce selezionata per i video AI è quella denominata "Patrick".

Prendendo in considerazione le operazioni eseguite per un video AI realizzato e relativo a un progetto aziendale, sono state inserite una serie di schermate provenienti dall'utilizzo dell'applicativo ElevenLabs.

Figura 5.33 Selezione voce denominata "Freya", dalle voci usate di recente nel tool. Subito dopo è presente la voce "Patrick". Ogni voce mostra, nelle etichette colorate, le proprie caratteristiche

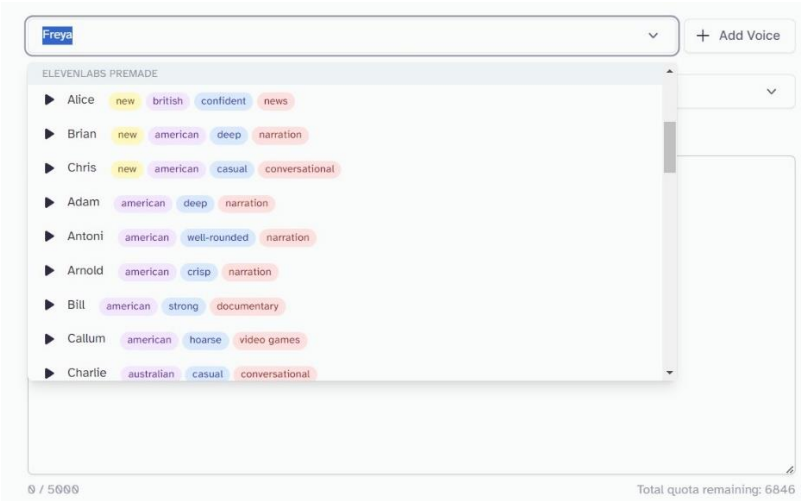

Figura 5.34 Elenco di alcune voci presenti nel catalogo del tool, tra cui la voce selezionata "Freya"

Inoltre, le voci generate, appartengono al catalogo di voci denominato come "Eleven Multilingual 2", che contiene 29 lingue integrate e utilizzabili nelle voci che si vogliono generare.

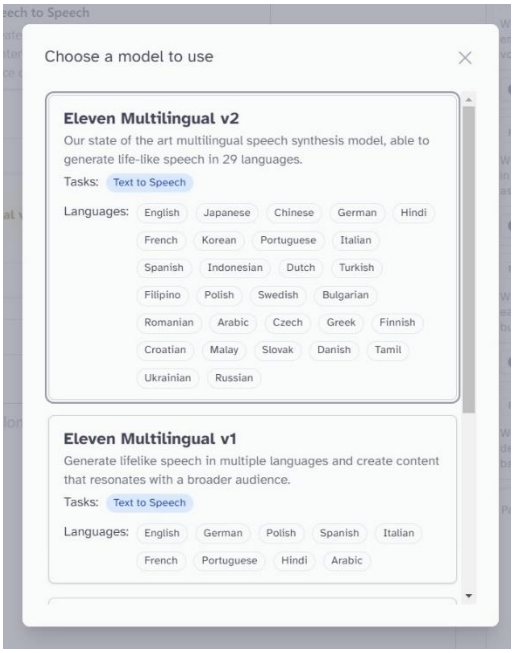

Figura 5.35 Catalogo di lingue che si possono utilizzare nelle voci da generare (sotto è mostrato un altro catalogo più datato)

Per le voci scelte, non sono state modificate le impostazioni vocali e sono state mantenute quelle di default. Se ne riporta, ugualmente, il contenuto offerto dall'applicativo, relativo alle voice settings.

Figura 5.36 Impostazioni vocali gestibili per la voce "Freya"

L'area di testo vuota su cui il candidato ha inserito gli script testuali dei progetti aziendali da generare in voce, è presente nella schermata che segue.

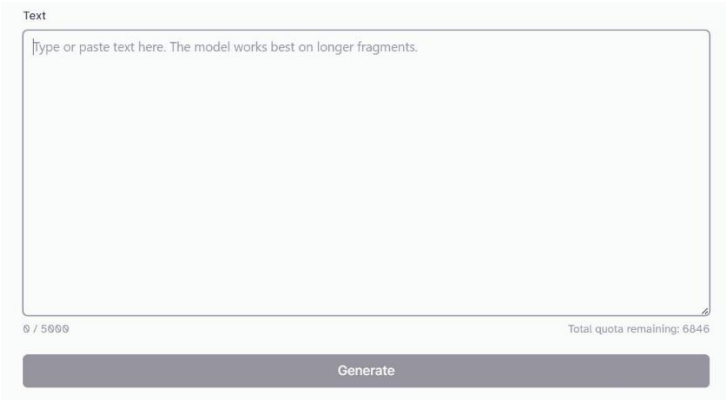

Figura 5.37 Area di testo vuota, da riempire con il prompt testuale che si desidera generare in voce

Nel caso del progetto considerato e video AI realizzato, è stata inserita un'altra schermata in cui si nota che, al centro, è presente il tasto "Generate" che consente di trasformare e generare lo script testuale inserito, in voce (di "Freya" in questo caso). Relativamente al testo inserito, questo è relativo a una scena video di un progetto aziendale trattato e, dunque, le tre righe aggiunte formano lo script testuale (generato nella fase 1) della scena video relativa.

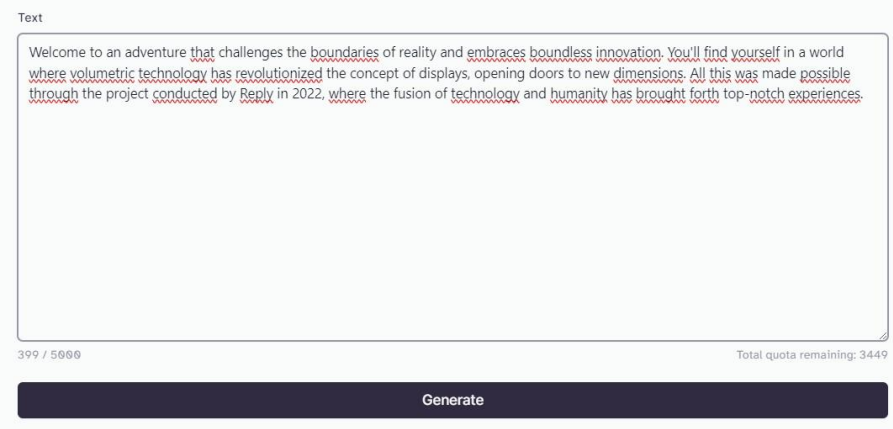

Figura 5.38 Area di testo con inserimento di uno script testuale, relativo a una scena video. Al centro è mostrato il bottone "Generate", che permette di generare lo script testuale in voce (di Freya in questo caso)

Dalle schermate mostrate, si intuisce come gli script testuali generati nella fase 1, relativi ai progetti aziendali considerati, sono stati generati in voce. In questo modo, sono state ottenute le voci narranti dei video AI realizzati.

5.2.5 Sottotitoli

Per quanto riguarda i video AI dei progetti aziendali, realizzati mediante le diverse fasi del workflow, una considerazione è utile anche per i sottotitoli, inseriti all'interno dei video.

Il candidato non ha applicato direttamente una fase di inserimento sottotitoli, né dall'uso del software di video editing adottato (Adobe Premiere) e neanche da qualche tool AI Generativo.

Ma sulla piattaforma video interna all'azienda Reply, sulla quale è avvenuta la pubblicazione dei contenuti audiovisivi AI generativi, è implementata l'opzione di attivare, oppure no, i sottotitoli per i video AI che, eventualmente, sono generati in automatico. I sottotitoli si presentano in lingua inglese, proprio come gli script creati e generati nella fase 1 del flusso di lavoro rispettato dal candidato.

5.2.6 Video editing finale

In quest'ultima fase di workflow, il candidato ha inserito le diverse clip video AI, ottenute in Runway nella fase 3, all'interno del software di video editing Adobe Premiere.

Il montaggio dei video AI finale ha come elemento principale l'inserimento in serie delle clip o scene video AI generate e prodotte.

I contenuti inseriti e gestiti in Premiere sono, dunque, stati post-prodotti sia lato video che lato audio, con la timeline composta da:

- Clip video AI;
- Tracce audio delle voci generate e relative agli script;
- Effetti sonori aggiuntivi per dare maggiore enfasi alla narrazione del progetto aziendale trattato in uno specifico video;

- Il format visivo pensato e creato dal candidato, che è comune e caratterizza i video AI realizzati;
- Le sigle aziendali, presenti all'inizio e alla fine dei video AI, utilizzate per indicare la creazione, da parte del candidato, di video episodi che faranno parte di una serie video originale, fruibile nella piattaforma video interna di Reply, interamente prodotta con l'AI Generativa.

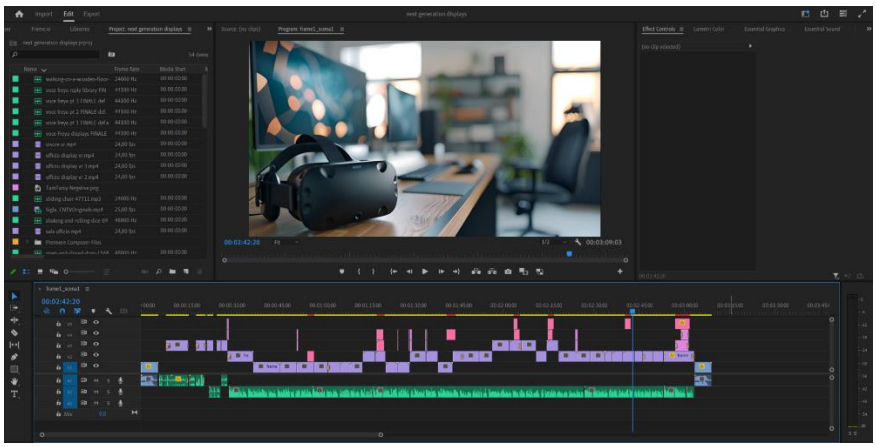

Figura 5.39 Timeline di un video AI realizzato e relativo a un progetto aziendale selezionato

In particolare, il format, ideato, creato e inserito visivamente dal candidato, consiste nella rappresentazione (a inizio e fine video AI) di un gioco da tavola in cui, il giocatore raffigurato, si sistema e si prepara a giocare e, una volta pronto, lancia dapprima un dado che mostra una cifra numerica e, dal numero ottenuto dal lancio, sposta la pedina per un certo numero di posizioni, fino a raggiungere la posizione desiderata. Al raggiungimento della posizione interessata, tramite spostamento della pedina da parte del giocatore, si esegue un movimento di camera di tipo "zoom-in" e inizia la parte di video AI contenente le scene narranti il progetto aziendale e si viene catapultati in una nuova dimensione, che è quella del progetto realizzato. Quindi, idealmente, il video AI che viene visto, è relativo alla posizione raggiunta nel gioco da tavola.

Per il successivo video AI realizzato, il format si ripete in maniera analoga, mostrando l'idea di aver raggiunto, dopo il lancio del dado, una nuova posizione del gioco, che segue quella raggiunta nel video AI precedente, a cui corrisponde un nuovo video AI. L'idea è stata quella di creare una sorta di percorso all'interno del gioco da tavola, fino a raggiungere l'ipotetico traguardo, che corrisponde all'ultimo video AI della serie originale.

Nelle successive immagini mostrate, sono raffigurati i 4 frame che appaiono animati nei video AI realizzati dal candidato e che rappresentano l'idea di format ideato (quello del gioco da tavola).

Figura 5.40 Frame format raffigurante il gioco da tavola

Figura 5.41 Frame format raffigurante il lancio del dado

Figura 5.42 Frame format raffigurante lo spostamento pedina

Figura 5.43 Frame format raffigurante la pedina posizionata e ferma

6. Conclusioni

6.1 Scenari futuri dell'AI Generativa

Dal presente lavoro di tesi condotto, e dalle ricerche effettuate, emerge come l'AI Generativa, nelle sue differenti sfaccettature, sta diventando sempre di più una tecnologia o un medium utile alla realizzazione di diverse tipologie di contenuti, soprattutto quelli legati alla Video e Image Generation, con immagini e video generabili in maniera affidabile e innovativa.

Ma più in generale, tra i diversi possibili scenari futuri e le ipotetiche evoluzioni che si possono presentare per l'AI Generativa, abbiamo:

- La personalizzazione e generazione estrema dei media (video, immagini, voci, audio): le tecnologie di AI generativa potrebbero consentire una personalizzazione e generazione estrema dei media, con sistemi e applicativi in grado di generare contenuti su misura per le preferenze individuali degli utenti, ad esempio prodotti audiovisivi e film personalizzati, esperienze di gioco uniche e storie narrative interattive;
- Gli assistenti virtuali creativi: questi potrebbero diventare più creativi e proattivi, in grado non solo di rispondere alle richieste degli utenti, ma anche di suggerire idee originali per nuove generazioni e creare, su richiesta, contenuti multimediali articolati come: disegni, musica o scritture estese di testi;
- La rivoluzione nell'industria dell'arte creativa: l'AI generativa potrebbe trasformare l'industria dell'arte e del design, consentendo agli artisti, ai designer e ai creatori di contenuti di sfruttare strumenti avanzati per esplorare nuove forme di espressione creativa e collaborare con algoritmi per generare opere d'arte innovative e moderne;
- Lo sviluppo di nuove forme di intrattenimento: potrebbero affermarsi e svilupparsi grazie all'AI generativa. Forme come esperienze immersive basate su realtà virtuale e aumentata, spettacoli teatrali interattivi e esperienze cinematografiche personalizzate.

Mentre, l'avanzamento della Generative AI nel contesto video potrebbe portare a una serie di sviluppi innovativi:

- Generazione di contenuti video sempre più realistici: gli algoritmi di AI generativa potrebbero diventare sempre più capaci di generare video realistici, compresi volti umani e ambienti, che potrebbero essere utilizzati in applicazioni come la produzione cinematografica, la pubblicità e la creazione di contenuti online;
- Produzione automatica di contenuti video: con l'AI generativa, potrebbe diventare più facile automatizzare il processo di produzione video, consentendo la generazione automatica di sequenze animate, effetti speciali e grafica in movimento per diversi contenuti finali come: film, programmi televisivi e video online;
- Personalizzazione dei video avanzata: le tecnologie di generative AI potrebbero consentire una personalizzazione avanzata dei video, con sistemi in grado di adattare il contenuto in tempo reale alle preferenze degli utenti, creando esperienze video altamente coinvolgenti e rilevanti per ciascun individuo;
- Ricostruzione e restauro di video storici: l'AI generativa potrebbe essere impiegata per ricostruire e restaurare video storici danneggiati o di bassa qualità, consentendo di preservare e valorizzare il patrimonio visivo dell'umanità in modo più accurato e dettagliato;
- Creazione di mondi virtuali ed effetti speciali: gli algoritmi di AI generativa potrebbero essere utilizzati per creare mondi virtuali realistici e dettagliati, utilizzati in settori come i videogiochi, la formazione virtuale e la simulazione di scenari reali per la ricerca scientifica e industriale e potrebbero essere utili per creare effetti speciali spettacolari e sofisticati per film, giochi e altre

applicazioni immersive, consentendo agli artisti e ai creatori di esplorare nuove frontiere della narrativa visiva.

In sintesi, il futuro della generative AI nel campo dei video promette di essere estremamente innovativo e rivoluzionario, con il potenziale di trasformare radicalmente la produzione e la fruizione dei contenuti video in tutto il mondo.

Dunque, l'AI Video Generation si troverà ad affrontare diverse sfide che la porteranno ad essere una delle categorie di Generative AI più diffuse e affidabili, diventando accessibile a chiunque e offrendo ai fruitori e creatori diversi servizi avanzati.

6.2 Future work

Relativamente all'AI Generativa, e all'ambito AI Video Generation, i lavori futuri potrebbero riguardare lo sviluppo di applicativi nuovi, avanzati e sofisticati che potrebbero fornire funzionalità migliori ed efficienti. Inoltre, i lavori e gli sviluppi futuri possono anche condurre gli utenti a nuovi approcci e nuove modalità di fruizione, per i rispettivi tool AI prodotti.

Tra i nuovi tool in fase di sviluppo è doveroso citare Sora, il nuovo tool di AI Video Generation implementato da OpenAI, che si è già stabilita come una delle aziende sviluppatrici migliori in ambito Generative AI, tramite la creazione di tool del calibro di ChatGPT e DALL-E 2 e 3.

Infine, la nascita di nuovi strumenti AI generativi (tra cui i tool di AI Video Generation) può spingere l'utente creatore ad adottare nuovi flussi di lavoro (o workflow) affidabili, dimostrandosi differenti rispetto al flusso di lavoro adottato dal candidato per svolgere una parte del progetto di tesi (i video AI realizzati e relativi ai progetti aziendali).

7. Ringraziamenti

E’ doveroso dedicare questo spazio della mia tesi a tutte le persone che mi hanno supportato nel mio percorso di crescita universitaria e professionale.

Ringrazio la mia famiglia, con i miei genitori Martino e Anna e mio fratello Davide che mi hanno sempre sostenuto in tutti i momenti, dagli esami affrontati ai giorni in cui tutto sembrava non andare per il verso giusto. Ringrazio i miei nonni Stefano, Concetta, Rosa e Domenico che mi hanno mostrato la loro vicinanza spronandomi ad affrontare tutto con consapevolezza e tenacia.

Ringrazio i miei amici di una vita e quelli conosciuti lungo l’intenso percorso universitario e non, per essermi stati d’aiuto nei momenti più difficili. Le ansie, le notti insonni, lo stress vissuto durante alcuni momenti di questo mio percorso sono stati alleviati e messi da parte anche grazie a voi.

Inoltre, un ringraziamento speciale lo rivolgo ai miei amici, colleghi e compagni di Reply, che in questi mesi mi hanno sostenuto lungo tutto il percorso collaborativo e di tesi, invogliandomi a dare sempre di più e a crescere e migliorare sia sotto l’aspetto umano che professionale.

8. Bibliografia e Sitografia

[1] https://tech4future.info/intelligenza-artificiale-cose-applicazioni/

[2] https://tech4future.info/generative-ai-cose-applicazioni/

[3] Haziqa Sajid – "*Articolo sui modelli di diffusione"* (2023): https://www.unite.ai/it/diffusion-models-in-ai-everything-you-need-to-know/

[4] https://arxiv.org/abs/2006.11239

[5] Jonathan Ho, Ajay Jain, Pieter Abbeel - "*Denoising Diffusion Probabilistic Models*" (2020): https://arxiv.org/pdf/2006.11239.pdf

[6] https://www.terrahunt.com/blog/ai-video-generation-advantages-disadvantages

[7] https://speechify.com/blog/what-is-the-future-of-ai-video-generation/?landing_url=https%3A%2F%2Fspeechify.com%2Fblog%2Fwhat-is-the-future-of-ai-video-generation%2F

[8] https://emu-video.metademolab.com/

[9] https://arxiv.org/abs/2311.10709

[10] GenAI, Meta "*EMU VIDEO: Factorizing Text-to-Video Generation by Explicit Image Conditioning*" (2023) https://arxiv.org/pdf/2311.10709.pdf

[11] https://medium.com/@dbhatt245/runway-gen-2-the-next-step-forward-for-generative-ai-an-introduction-b85bc90d3e45

[12] https://www.simplilearn.com/keras-vs-tensorflow-vs-pytorch-article#:~:text=For%20extensive%20projects%20with%20significant,PyTorch%20is%20the%20suitable%20option

[13] *Spiegazione illustrata dei framework ML di Runway:* https://www.youtube.com/watch?v=4L86D_fU6sQ

[14] A*pproccio per la generazione di script con chatGPT:* https://www.youtube.com/watch?v=kbBGyXr-U2E