



**Politecnico
di Torino**

Politecnico di Torino

**Laurea Magistrale in
Ingegneria del Cinema e dei Mezzi di Comunicazione**

A.A. 2024/2025

Sessione di Laurea Luglio 2025

IA e Arte

Narrare al pubblico l'impatto emotivo delle performance musicali attraverso
video generati con strumenti di intelligenza artificiale generativa

Relatori:

Prof. Andrea BOTTINO
Prof. Tatiana MAZALI

Candidato:

Simone BOTOSSO

Abstract

Le tecnologie di Intelligenza Artificiale (IA) stanno influenzando sempre più i processi creativi, in particolare nell'ambito audiovisivo. Questa tesi indaga le potenzialità espressive e la fattibilità produttiva di contenuti audiovisivi generati tramite IA, a partire dalle emozioni raccolte durante performance musicali dal vivo. Il progetto è stato sviluppato in collaborazione con Narrazioni Parallele, iniziativa culturale triennale che esplora la relazione tra Intelligenza Naturale e Artificiale in contesti musicali. I dati raccolti durante due eventi dell'iniziativa hanno costituito la base per questo studio sperimentale.

La ricerca si apre con un'analisi dello stato dell'arte del rapporto tra IA e arti visive, con un focus sul settore cinematografico. Viene poi effettuata una ricognizione delle principali tecnologie generative IA nei campi della scrittura creativa, sintesi di immagini, generazione di video, di suoni, di voci e di musiche, al fine di individuare gli strumenti più efficaci per ciascuna fase del processo produttivo.

Il workflow sperimentale si articola in quattro fasi principali: strutturazione dei dati, generazione automatica dello script, definizione visiva tramite storyboard e produzione delle clip mediante tecniche Image-to-Image Video e Image-to-Video. I materiali audiovisivi sono stati completati con suoni, voci e musiche generate da IA, portando alla fase finale di montaggio e proiezione pubblica. I feedback del pubblico sono stati raccolti e analizzati per valutarne l'impatto emotivo e la coerenza narrativa.

I risultati evidenziano il potenziale espressivo di questi strumenti, ma anche i limiti attuali in termini di controllo e precisione produttiva. Tuttavia, l'evoluzione rapida di queste tecnologie ne suggerisce una probabile integrazione futura all'interno di pratiche creative sperimentali e ibride.

Indice

1	Introduzione	6
1.1	Panoramica della Tesi	6
1.2	Obiettivi della Tesi	6
1.3	Presentazione del Progetto “Narrazioni Parallele”	7
1.3.1	Evento: Una Passeggiata Creativa	7
1.3.2	Evento: Song Writing, Intelligenza Naturale Generativa	7
2	Stato dell’Arte	8
2.1	Come l’IA sta influenzando il mondo dell’arte	9
2.1.1	IA nel Cinema	11
2.2	IA per la Scrittura Creativa	16
2.2.1	Funzionamento dei Large Language Models	16
2.2.2	Comparazione IA per la Generazione Testuale	21
2.3	IA per la Generazione di Immagini	24
2.3.1	Funzionamento delle IA per la Generazione di Immagini	24
2.3.2	Confronto IA per la Generazione di Immagini	30
2.4	IA per la Generazione di Video	33
2.4.1	Funzionamento delle IA per la Generazione di Video	33
2.4.2	Framework Applicativo: ComfyUI	37
2.4.3	Comparazione IA per la Generazione di Video	39
2.5	IA per la Generazione di Musica, Suoni e Voci	44
2.5.1	Funzionamento dei Modelli per la Generazione Musicale	44
2.5.2	Funzionamento dei Modelli per la Generazione di Effetti Sonori	45
2.5.3	Funzionamento dei Modelli per la Generazione Vocale	46
2.5.4	Comparazione IA per la Generazione Musicale	46
2.5.5	Comparazione IA per la Generazione di Effetti Sonori	48
2.5.6	Comparazione IA per la Generazione Vocale	50
3	Elaborato di tesi	53
3.1	Tipologia di Workflow	54
3.1.1	Approccio Online	54
3.1.2	Approccio Locale	56
3.1.3	Scelta del Workflow	58
3.2	Analisi dei Dati	59
3.2.1	Resoconto Survey “Una Passeggiata Creativa”	59
3.2.2	Resoconto Survey “Song Writing Intelligenza Generativa”	61
3.2.3	Analisi dei Dati Ottenuti	64

3.3	Generazione dello Script	66
3.3.1	Definizione del Prompt	66
3.3.2	DeepSeek R1	67
3.3.3	Claude 3.7 Sonnet	69
3.3.4	ChatGPT-4o (Feb 2025)	72
3.3.5	Valutazione e Scelta dello Script	74
3.4	Generazione dello Storyboard	76
3.4.1	Scelta dello stile visivo	76
3.4.2	Il Fenomeno dell’“Uncanny Valley”	77
3.4.3	Espressionismo Come Amplificatore di Emozioni	79
3.4.4	Confronto tra i Modelli Scelti	82
3.4.5	Confronto Output e Scelta del Modello	85
3.4.6	Generazione dello Storyboard tramite Recraft V3	87
3.5	Generazione dei Video	91
3.5.1	Generazione dei Video Tramite Adobe Firefly	91
3.6	Generazione Musica, Suoni e Voiceover	96
3.6.1	Generazione della musica	96
3.6.2	Generazione dei Suoni e del Voiceover	97
3.7	Montaggio del Video	100
3.8	Valutazione delle Risposte del Pubblico	103
	Conclusioni	107
3.9	Sviluppi futuri	108
	Ringraziamenti	110
	Bibliografia	111

Capitolo 1

Introduzione

1.1 Panoramica della Tesi

Questa tesi mira ad esplorare il punto di incontro tra l'arte tradizionale e le moderne tecnologie di Intelligenza Artificiale (IA), con l'obiettivo di creare un video capace di suscitare emozioni analoghe a quelle evocate da un concerto dal vivo.

Il progetto si è sviluppato tramite gli eventi di “Narrazioni Parallele”, una serie di concerti durante i quali è stato raccolto un ampio spettro di dati relativi alle reazioni emotive dei partecipanti. Questi dati sono stati quindi catalogati e analizzati per identificare pattern utili alla generazione di contenuti video.

Successivamente, è stato esaminato il potenziale delle intelligenze artificiali per automatizzare ed innovare il tradizionale processo di produzione audiovisiva. Sono state ricercate e testate le migliori soluzioni attualmente presenti sul mercato per ogni fase della realizzazione di un video: dalla scrittura dello script alla realizzazione dello storyboard, fino alla creazione dei contenuti multimediali e al montaggio finale. Sono state coinvolte molteplici aree di generazione di contenuti: non solo per la realizzazione dei video veri e propri, ma anche sistemi per generare testi, immagini e musiche.

Una volta identificati i migliori tools per la ricerca, si è compiuta la realizzazione del video finale. A seguito della proiezione del prodotto, si è analizzato il feedback fornito dal pubblico.

Al fianco della parte più pratica della tesi, si accompagna un lavoro di ricerca e analisi dei vari strumenti e si approfondisce la logica adottata dall'IA per la generazione di contenuti.

1.2 Obiettivi della Tesi

Gli obiettivi che si vogliono raggiungere con la seguente tesi riguardano:

- Trovare il metodo più efficace di raccolta dati durante e dopo gli eventi, coinvolgendo le persone (performer e pubblico) con tecnologie adatte a renderle partecipi del processo.
- Raccogliere e categorizzare i dati che forniranno la base per i prompt per la generazione del video con software IA.

- Utilizzare la IA generativa mappando nei contenuti eventuali stereotipizzazioni discriminanti con un approccio aperto alle pluralità (di generi, abilità, culture, ecc.).
- Realizzare un video creativo con IA partendo da una “sceneggiatura” e un prompt costituiti dagli input del pubblico presente ai concerti del progetto Narrazioni Parallele, generando un artefatto artistico con una propria autonomia estetica.
- Validare il prodotto finale attraverso una verifica qualitativa, coinvolgendo gli utenti nella valutazione dell’efficacia espressiva del video generato.

1.3 Presentazione del Progetto “Narrazioni Parallele”

Narrazioni Parallele è un progetto culturale triennale, avviato nel 2024 a Torino, che si propone di esplorare l’interazione tra intelligenza naturale e intelligenza artificiale nel contesto musicale. Promosso dall’associazione Lucas in collaborazione con l’Accademia Corale Stefano Tempia, Cultura e Salute APS, Filarmonica TRT (Teatro Regio Torino), Biblioteche Civiche Torinesi e il Politecnico di Torino, il progetto si articola in due direttrici principali:

- **Narrazione-Partecipazione:** Questa sezione prevede l’organizzazione di concerti e workshop interattivi all’interno delle Biblioteche Civiche Torinesi, con l’obiettivo di coinvolgere attivamente un pubblico intergenerazionale. Durante questi eventi, i partecipanti sono invitati a condividere le proprie emozioni e percezioni suscitate dall’ascolto musicale, contribuendo così alla co-creazione artistica.
- **Narrazione-Contemporaneità:** Questo percorso offre a giovani compositori under 35 l’opportunità di presentare le proprie opere attraverso “call for composers”. Le composizioni selezionate vengono sviluppate con il supporto di mentori e presentate in concerti che combinano musica classica ed elettronica, esplorando nuovi linguaggi e tecnologie.

In questo contesto, l’intelligenza artificiale funge da elemento di connessione tra queste due dimensioni.

1.3.1 Evento: Una Passeggiata Creativa

Svoltosi il 30 ottobre 2024 presso la Biblioteca Civica Italo Calvino, questo evento ha visto l’esibizione del Quintetto Prestige della Filarmonica TRT, che ha eseguito una trascrizione per quintetto di fiati di “Quadri di un’esposizione” di Modest Mussorgsky, arrangiata da Joachim Linckelmann. Durante e dopo l’esecuzione musicale, il pubblico è stato incoraggiato a condividere le emozioni suscitate dai vari “Quadri” tramite sondaggi.

1.3.2 Evento: Song Writing, Intelligenza Naturale Generativa

Tenutosi il 22 novembre 2024, sempre presso la Biblioteca Civica Italo Calvino, questo workshop musicale è stato condotto dalla cantante Arianna Gallo, accompagnata dal chitarrista Luca Ammazzini. L’evento ha coinvolto attivamente il pubblico nella co-creazione di una canzone. Come nell’evento precedente, anche in questo caso al pubblico è stato richiesto di condividere le proprie emozioni tramite un sondaggio.

Capitolo 2

Stato dell'Arte

Negli ultimi anni, l'intelligenza artificiale ha vissuto un'accelerazione senza precedenti, modificando profondamente numerosi settori dell'attività umana. Tra questi, il mondo dell'arte e della produzione audiovisiva ha sperimentato una trasformazione radicale, nella quale l'IA non è più soltanto uno strumento tecnico, ma un vero e proprio agente creativo e produttivo.

Questo capitolo si propone di analizzare in che modo i modelli generativi – in particolare i Large Language Models (LLM) ed altre varie categorie di modelli per la generazione di immagini, video, musica, effetti sonori e voce – stiano ridefinendo i processi di creazione artistica e cinematografica. Attraverso un'analisi dettagliata dello stato dell'arte e dei benchmark più aggiornati, il lavoro intende fornire una panoramica critica delle potenzialità, delle limitazioni e delle implicazioni culturali dell'intelligenza artificiale applicata alla narrazione e all'estetica visiva.

Verranno esplorate quattro aree principali: la scrittura creativa automatizzata, la generazione di immagini, di video e di suono (che include la generazione di musica, di effetti sonori e di sintesi vocale). Per ciascuna di queste verranno approfonditi i principi di funzionamento dei modelli e i progressi recenti.

2.1 Come l'IA sta influenzando il mondo dell'arte

La rapida evoluzione dell'IA degli ultimi anni, spinta dai progressi di ricerca nel campo del machine learning, ha avuto come conseguenza la ricerca di nuovi metodi di utilizzo di questa nei campi più disparati.

Tra questi vi è l'applicazione dell'IA come strumento artistico: che sia per la generazione di immagini, di video o di musica, al giorno d'oggi è pratica comune avvalersene per i più disparati motivi. L'applicazione di questa tecnologia nel campo dell'arte ha creato non pochi dubbi, in quanto è pensiero comune che l'arte e la sua comprensione siano capacità ancora proprie del solo essere umano e della sua creatività. Eppure, si è assistito ad un incremento di interesse da parte di artisti e ricercatori sull'esplorazione del potenziale creativo dell'intelligenza artificiale.

“*Creativity can be defined as the ability to generate novel, and valuable, ideas.*” così Margaret A. Boden definisce il concetto di creatività nel suo articolo “Computer Models of Creativity”¹, in cui ne vengono distinti tre diversi tipi che l'IA può manifestare:

- Combinatoria, ovvero la combinazione di materiale pre-esistente in nuovi modi, creando associazioni tra stili e tematiche. Questo tipo di approccio è esemplificato da progetti come “The Next Rembrandt”², del 2016, in cui l'IA ha analizzato le opere del maestro olandese per creare un nuovo dipinto che ne emulasse lo stile e la tecnica.
- Esplorativa, ovvero l'esplorazione di uno stile o di un paradigma già esistente. Artisti come Refik Anadol utilizzano queste reti per esplorare le possibilità all'interno di un dominio stabilito. Il suo progetto “Machine Hallucinations”³, del 2021, esplora le potenzialità dell'architettura neurale per creare paesaggi onirici in movimento, rimanendo all'interno dei parametri estetici della fluidità visiva contemporanea.
- Trasformativa, ovvero l'alterazione delle regole che definiscono un dominio, e da cui si produce qualcosa di nuovo. Quest'ultima, considerata la forma più radicale di creatività, rimane principalmente appannaggio dell'ingegno umano, ma l'IA sta rapidamente avanzando nell'ambito della creatività esplorativa, producendo opere che sorprendono persino i loro creatori. Alcuni esperimenti come quelli del sistema “AICAN” sviluppato da Ahmed Elgammal presso la Rutgers University mostrano tentativi di superare i confini stilistici esistenti, generando opere che non appartengono a nessuna corrente artistica precedente⁴.

Seppur vi possano essere dubbi se l'output dell'IA possa effettivamente essere considerata arte o meno, è innegabile come questa tecnologia venga usata e possa essere uno strumento utile. Principalmente vi sono due ambiti di ricerca legati al binomio “Arte e AI”⁵:

- Analisi dell'arte esistente: si sfrutta l'IA per studiare, catalogare e interpretare il patrimonio artistico. È emblematico, in questo contesto, il caso della piattaforma

¹Margaret A. Boden, *Computer Models of Creativity*, 2009. [1]

²J. Walter Thompson Amsterdam, *The Next Rembrandt: Recreating the work of a master with IA*, 2016. [2]

³Refik Anadol, *Machine Hallucinations: Nature Dreams*, 2021. [3]

⁴Ahmed Elgammal et al., *CAN: Creative Adversarial Networks, Generating “Art” by Learning About Styles and Deviating from Style Norms*, 2017. [4]

⁵Eva Cetinic and James She, *Understanding and Creating Art with AI: Review and Outlook*, 2021. [5]

“Art Recognition”⁶ che, sviluppata da un team svizzero, impiega reti neurali convoluzionali per l’autenticazione di opere d’arte, raggiungendo un’elevata percentuale di accuratezza nell’identificazione di falsi, superando in alcuni casi l’occhio esperto dei curatori.

- Creazione di nuova arte: artisti come Mario Klingemann hanno integrato l’IA nei loro processi creativi. La sua opera “Memories of Passersby I”⁷, del 2018, utilizza reti neurali per generare ritratti in tempo reale, creando un flusso infinito di volti umani che non sono mai esistiti.

Le IA hanno, inoltre democratizzato l’accesso alla creazione visiva, permettendo anche a non-artisti di generare immagini complesse a partire da descrizioni testuali. Questo ha dato vita a un nuovo genere di “prompt artists”, creativi specializzati nella formulazione precisa di istruzioni testuali per ottenere risultati visivi specifici.

Tutto ciò è stato reso possibile dal processo di digitalizzazione delle opere d’arte degli ultimi decenni, che ha permesso un notevole incremento dei dati a cui far attingere le diverse IA in fase di addestramento: la disponibilità di dataset ampi e ben annotati è un requisito fondamentale per l’adozione di modelli di deep learning in questi ambiti.

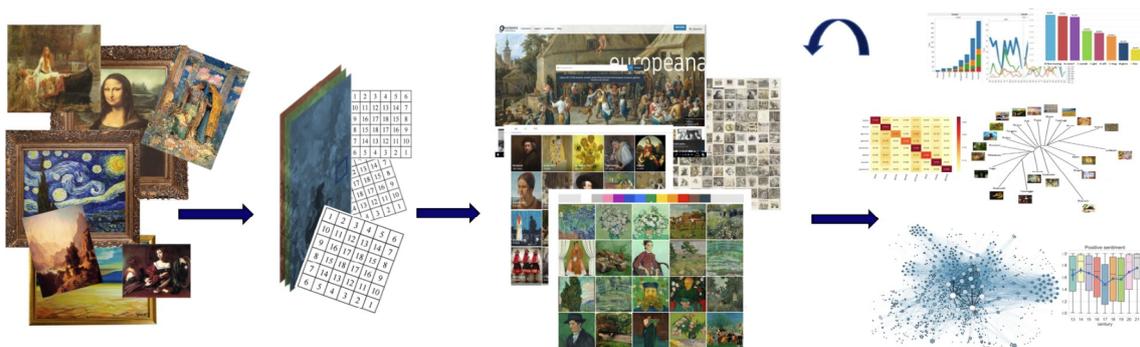


Figura 2.1: Illustrazione del processo dalla digitalizzazione all’analisi computazionale avanzata.

Attualmente, il rapporto IA - Artista è mutato: l’intelligenza artificiale, da mero strumento, è diventato un vero e proprio assistente. Come evidenziato da Aaron Hertzmann in “Visual Indeterminacy in GAN Art”⁸, questa transizione ha generato un nuovo paradigma di co-creazione dove l’artista umano e l’intelligenza artificiale lavorano in simbiosi. L’umano fornisce direzione creativa, cura e contesto, mentre l’IA contribuisce con capacità generative e associative uniche.

Helena Sarin, artista visiva con un background nell’ingegneria informatica, ha sviluppato un approccio che definisce “collaborativo”, in cui addestra GAN (Generative Adversarial Networks) con i propri disegni e dipinti, creando un sistema che genera variazioni uniche del suo stile personale, che poi lei seleziona e raffina ulteriormente⁹.

⁶Art Recognition, *Art Recognition: AI-based Art Authentication*, 2021. [6]

⁷Mario Klingemann, *Memories of Passersby I*, 2018. [7]

⁸Aaron Hertzmann, *Visual Indeterminacy in Generative Neural Art*, 2020. [8]

⁹Kate Vass, *Helena Sarin: Neural Bricolage*, 2024. [9]

Da ottobre 2018, quando l'opera "Portrait of Edmond Belamy" creata dal collettivo Obvious venne venduta da Christie's per 432.500 dollari¹⁰, si è manifestato un interesse crescente verso l'arte prodotta dall'intelligenza artificiale, accompagnato dalla necessità di esaminare gli aspetti fondamentali di questo nuovo movimento nel panorama artistico contemporaneo.



Figura 2.2: Portrait of Edmond Belamy.

Il caso del "Portrait of Edmond Belamy" ha suscitato in particolare la discussione sull'autorialità e sull'etica: quando un'opera è generata da un sistema di intelligenza artificiale, chi può essere considerato il suo vero autore? È l'artista umano che ha concepito l'idea e fornito l'input testuale o visivo (il cosiddetto prompt)? È lo sviluppatore che ha ideato e addestrato l'algoritmo? Oppure, paradossalmente, è l'IA stessa, in quanto autrice materiale dell'output?

Questa ambiguità sfida le definizioni tradizionali di autorialità e creatività, che storicamente presuppongono l'intenzionalità e la coscienza dell'artista. Le IA, invece, generano contenuti basandosi su correlazioni matematiche nei dati, senza alcuna comprensione semantica o volontà espressiva. Eppure, le opere risultanti possono evocare emozioni, raccontare storie e perfino influenzare la cultura popolare, complicando ulteriormente la questione.

Dal punto di vista legale, il dibattito è ancora aperto. In molti ordinamenti giuridici, i diritti d'autore sono riconosciuti solo a persone fisiche. Tuttavia, casi recenti — come la rimozione di immagini generate con IA da concorsi artistici — hanno mostrato la necessità di ridefinire le regole. Alcune giurisdizioni, come quella degli Stati Uniti, hanno già dichiarato che le opere generate autonomamente da IA non possono essere coperte da copyright, mentre altre stanno valutando l'adozione di nuovi strumenti normativi che contemplino la co-creazione uomo-macchina.

2.1.1 IA nel Cinema

Fin dalle sue origini, l'industria cinematografica è sempre stata influenzata dalle innovazioni tecnologiche. Basti pensare ai suoi albori, con l'introduzione del cinematografo, o alla rivoluzione del sonoro degli anni '30, o ancora all'avvento del colore negli anni '50.

¹⁰Christie's, *Is artificial intelligence set to become art's next medium?*, 2018. [10]

Tra queste rivoluzioni, quella a cui più si avvicina, rispetto all'IA, è l'avvento del digitale e della computer grafica all'inizio degli anni '90¹¹, che ha permesso di creare effetti speciali e mondi virtuali prima impensabili. Oggi l'IA sta ulteriormente trasformando il panorama cinematografico, introducendo nuove possibilità per ogni aspetto della produzione, dalla sceneggiatura alla post-produzione.

IA nella pre-produzione cinematografica

Nel 2024 è stato rilasciato il film “*The Last Screenwriter*”, diretto da Peter Luisi¹², il primo lungometraggio scritto da un'intelligenza artificiale, in particolare dal modello di linguaggio GPT-3.5. Il film racconta di uno sceneggiatore che inizia ad utilizzare un'IA di scrittura creativa, prima come assistente e poi come vero e proprio co-autore. È interessante notare che il messaggio che il film vuole trasmettere è che non bisogna dimenticarsi che le macchine non possono replicare correttamente l'esperienza umana e che “il tocco umano è insostituibile”.

Il sopracitato film è, però, un'estremizzazione di quello che è l'effettivo utilizzo delle IA nel mondo del cinema del giorno d'oggi: attualmente questi sistemi sono usati come assistenti o come metodi automatici per compiere specifiche operazioni.

Sempre per quanto riguarda la pre-produzione cinematografica, esistono diversi strumenti che intendono sostituire i processi decisionali soggettivi svolti dai dirigenti delle case di produzione e distribuzione¹³:

- *ScriptBook*: piattaforma online che utilizza tecniche di data mining, machine learning ed elaborazione del linguaggio naturale (NLP) per elaborare l'input (sceneggiature cinematografiche) e generare un output unico (analisi delle caratteristiche della sceneggiatura, della sua redditività commerciale sotto vari aspetti e una raccomandazione finale di approvazione o rifiuto), riducendo ciò che normalmente richiederebbe giorni o settimane di ricerca e di riflessioni a una questione di pochi minuti.
- *Cinelytic*: azienda IA che fornisce strumenti di analisi predittiva per supportare le decisioni nella fase di approvazione dei progetti. Nel 2020, Cinelytic ha annunciato una partnership con Warner Bros. per integrare la sua piattaforma di intelligenza artificiale nella loro pipeline di produzione cinematografica, consentendo una valutazione più rapida e accurata dei progetti. La Warner Bros ha affermato che non utilizzerà il sistema per automatizzare le decisioni su quali film realizzare, ma per ridurre il tempo dedicato a compiti ripetitivi come l'analisi dei costi legati al packaging, marketing e distribuzione, incluse le date di uscita.

IA nella produzione cinematografica

Durante la produzione dei contenuti, l'IA può essere utilizzata per migliorare gli aspetti tecnici e creativi del processo cinematografico. Nel 2019 è stata rilasciata da Disney la serie

¹¹Davide Tromba, 27 maggio 2025, *AI Storytelling and Cinematography: come i modelli open source stanno rivoluzionando la pipeline cinematografica.*, Conferenza presentata al Politecnico di Torino, Torino, Italia. [11]

¹²Peter Luisi, *The Last Screenwriter*, 2024. [12]

¹³Pei Sze Chow, *Ghost in the (Hollywood) machine: Emergent applications of artificial intelligence in the film industry*, 2020. [13]

TV “*The Mandalorian*”, che ha utilizzato un sistema di intelligenza artificiale chiamato “StageCraft”¹⁴ per creare ambienti virtuali in tempo reale, permettendo agli attori di recitare in scenari digitali senza dover ricorrere a green screen o set fisici complessi. Questo ha ridotto i costi e i tempi di produzione, consentendo una maggiore flessibilità creativa e garantendo un’illuminazione più coerente degli oggetti sul set.



Figura 2.3: Sul set di *The Mandalorian*

IA nella post-produzione cinematografica

Anche in fase di post-produzione, l’IA sta rivoluzionando il mondo cinematografico.

Emblematico è stato il caso di “*The Brutalist*”¹⁵, film del 2024 diretto da Brady Corbet e vincitore di tre premi Oscar, tra cui quello per il miglior film. Durante la post-produzione del film, tramite il software di intelligenza artificiale Respeecher, si è andati a modificare l’accento dei protagonisti al fine di migliorare la pronuncia di determinate lettere o dittonghi, per rendere più veritiero l’accento ungherese. Inoltre, si è fatto uso dell’IA generativa anche per i rendering “anni ‘80” di alcuni progetti visibili durante le scene finali. L’utilizzo dell’IA in un film di questo calibro ha suscitato un acceso dibattito, durante il quale il regista ha difeso l’uso dell’intelligenza artificiale generativa definendo il processo come “interamente manuale” e fatto “col massimo rispetto” per il lavoro degli attori.

Un’altra tipologia di ausilio dell’IA nella post-produzione è il ringiovanimento digitale degli attori, come si è visto nel film “*The Irishman*”¹⁶ di Martin Scorsese, dove l’IA è stata utilizzata per ringiovanire gli attori Robert De Niro, Al Pacino e Joe Pesci, permettendo loro di interpretare i personaggi in diverse fasi della vita senza dover ricorrere a controfigure o trucco pesante.

Infine, l’IA sta influenzando anche la distribuzione e il marketing dei film. Piattaforme come Netflix e Amazon Prime Video utilizzano algoritmi di raccomandazione basati su IA per suggerire contenuti agli utenti, analizzando le loro preferenze e i loro comportamenti di

¹⁴Troy Yarter, *The Volume: How “The Mandalorian” Revolutionized Filmmaking*, 2024. [14]

¹⁵Bethy Squires, *The Brutalist’s AI Controversy, Explained*, 2025. [15]

¹⁶Chris Lee, *How ‘The Irishman’ Used CGI & Special Effects on Actors*, 2020. [16]

visione. Questo ha portato a una maggiore personalizzazione dell'esperienza cinematografica, consentendo agli spettatori di scoprire film e serie TV che altrimenti potrebbero passare inosservati.

Sviluppi futuri

In tutta la filiera cinematografica non si è ancora visto un utilizzo massiccio dell'IA nel campo della *Video Generation*. Questo è dovuto al fatto che la generazione video per mezzo delle intelligenze artificiale presenta alcune problematiche significative¹⁷:

- Valutazione della recitazione e animazione: nella produzione IA attuale mancano strumenti per valutare efficacemente la qualità dell'animazione e della recitazione generata.
- Misurazione dell'impatto emotivo: è difficile quantificare e calibrare il reale impatto emotivo che il contenuto generato avrà sugli spettatori.
- Correzione iterativa: mancano meccanismi efficaci per correggere in modo mirato e progressivo gli elementi emblematici delle creazioni IA.
- Validazione degli elementi: è difficile mettere un fermo su un elemento e capire quando questo può essere considerato "soddisfacente" o meno. Inoltre, la mancanza di "semielaborati" validabili complica il processo creativo.

È proprio per queste difficoltà nel controllo qualitativo che si sono sviluppati processi produttivi e pipeline di lavorazione ibride, che integrano l'IA in diversi momenti del workflow e non in tutta la lavorazione.

È stato proposto un indice per valutare il livello di IA utilizzata in un film, chiamato *Human Control Index* (HCI):

$$\text{Human Control Index (HCI)} = \frac{\text{Potere Decisionale del Regista}}{\text{Autonomia dell'AI}}$$

Quando il risultato è approssimativamente infinito, significa che l'opera è interamente controllata dall'umano; se invece è approssimativamente zero, l'opera è interamente realizzata dall'IA. Tuttavia, questo approccio presenta significative limitazioni pratiche: la quantificazione del "potere decisionale" e dell'"autonomia dell'IA" rimane altamente soggettiva e dipendente dal contesto specifico di produzione. Inoltre, la formula non tiene conto della qualità dell'intervento umano o della complessità delle decisioni delegate all'IA. Per esempio, un regista che mantiene il controllo creativo su aspetti narrativi fondamentali ma delega all'IA la generazione di elementi visivi secondari potrebbe avere un HCI simile a quello di un collega che controlla ogni singolo frame ma utilizza l'IA per compiti puramente tecnici, pur rappresentando approcci creativi completamente diversi.

Una delle pipeline più promettenti è quella che prevede l'utilizzo di modelli IA utilizzati e addestrati in locale, tipologia di workflow che verrà approfondita nel paragrafo 3.1.2. Le

¹⁷Davide Tromba, 27 maggio 2025. *AI Storytelling and Cinematography: come i modelli open source stanno rivoluzionando la pipeline cinematografica.*, Conferenza presentata al Politecnico di Torino, Torino, Italia. [11]

problematiche principali di questa tipologia di lavorazione risultano essere l'hardware e le risorse richieste: attualmente sono necessari server con GPU di ultima generazione per poter eseguire questi modelli di IA generativa in modo efficiente, andando contro a quel senso di "democratizzazione" che l'IA dovrebbe portare con sé.

Negli ultimi anni, l'integrazione dell'intelligenza artificiale nella produzione cinematografica ha dato origine ad una nuova serie di concorsi e festival dedicati ai cortometraggi realizzati con l'ausilio di quest'ultima. Questi eventi sono nati per celebrare l'evoluzione del linguaggio audiovisivo e delle nuove forme di narrazione, e non solo l'innovazione tecnica.

In Italia si tengono annualmente diversi festival dedicati a questo tema:

- *Reply AI Film Festival*¹⁸: festival internazionale dedicato ai cortometraggi realizzati con IA e che offre la possibilità a tutti coloro che volessero produrre un proprio contenuto di partecipare gratuitamente. Ad ogni edizione partecipano artisti provenienti da tutto il mondo, evidenziando l'importanza internazionale di questo festival.
- *AI.motion*¹⁹: concorso indetto dall'università IULM di Milano in collaborazione con IULM AI Lab e la Neural Network Association. Il festival è dedicato a cortometraggi, video clip e spot pubblicitari creati con l'intelligenza artificiale, con una sezione dedicata ai personaggi digitali basati sull'intelligenza artificiale e agli avatar iperrealistici.

A livello internazionale, invece, si evidenziano eventi come:

- *AIFF 2025 – AI Film Festival*²⁰: introdotto da Runway AI, nota piattaforma online di video-generation, è uno dei più importanti festival dedicati a questa tematica.
- *AI International Film Festival*²¹: festival che si tiene a Hollywood mensilmente, con l'obiettivo di essere al passo con il rapido sviluppo dell'IA.

¹⁸Reply, *Reply AI Film Festival*, 2024. [17]

¹⁹IULM university, *AI.motion*, 2024. [18]

²⁰Runway AI, *AIFF 2025 – AI Film Festival*, 2025. [19]

²¹AI Film Fest, *AI International Film Festival*, 2025. [20]

2.2 IA per la Scrittura Creativa

Negli ultimi anni, i Large Language Models (LLM) hanno rivoluzionato il panorama dell'intelligenza artificiale applicata al linguaggio naturale, dimostrando capacità sempre più sofisticate nella generazione autonoma di contenuti testuali. Tra i diversi ambiti di applicazione, la scrittura creativa rappresenta una delle sfide più complesse e affascinanti, in quanto richiede non solo la competenza linguistica, ma anche la capacità di strutturare narrazioni coerenti, coinvolgenti e originali.

In questo contesto, la valutazione comparativa delle prestazioni dei LLM nella produzione di testi narrativi è diventata un terreno fertile per l'indagine critica e sperimentale. I modelli non vengono più giudicati unicamente sulla base della correttezza sintattica o grammaticale, ma anche sulla loro abilità nel gestire elementi narrativi strutturali come personaggi, ambientazioni, obiettivi e tono, integrandoli armonicamente in una storia completa.

Il presente paragrafo, dopo aver approfondito il funzionamento di questi sistemi, si concentra sull'analisi di un benchmark specificamente progettato per testare le capacità narrative dei LLM, ponendo particolare enfasi sulla loro abilità di incorporare dieci elementi narrativi fondamentali nella produzione di brevi racconti originali. Questo benchmark, somministrato a 32 modelli linguistici e valutato da una commissione di sei LLM distinti, consente di ottenere una visione comparativa dettagliata delle performance di ciascun modello, sia in termini di rispetto dei vincoli narrativi che di qualità stilistica e strutturale del testo prodotto.

2.2.1 Funzionamento dei Large Language Models

I Large Language Models (LLM) sono modelli di intelligenza artificiale progettati per analizzare, comprendere e generare linguaggio naturale. La loro efficacia deriva dalla combinazione di un'enorme quantità di dati testuali utilizzati durante l'addestramento con architetture neurali avanzate in grado di cogliere le strutture sintattiche, semantiche e pragmatiche del linguaggio.

Alla base di questi modelli vi sono reti neurali ispirate al funzionamento del cervello umano: in una rappresentazione semplificata, un neurone riceve segnali da altri neuroni e produce un segnale in uscita. L'utilizzo di una rete neurale prevede la fornitura di dati in input e l'osservazione del risultato in output: tra i due vi è una rete di neuroni. Il modo in cui i dati vengono modificati dalla rete viene condizionato dai "pesi", che regolano quanto l'informazione corrente debba influire: più il peso è alto, maggiore sarà l'influenza di quest'informazione sul risultato in output.

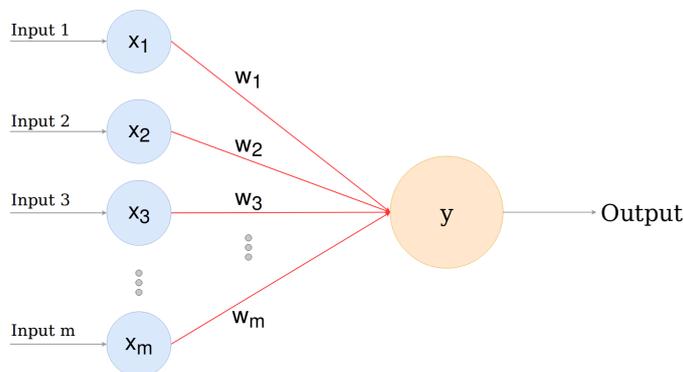


Figura 2.4: Struttura di un perceptrone (rete neurale semplice).

Questa struttura è regolata dalla seguente formula:

$$Output = y \left(b + \sum_{i=1}^n w_i x_i \right)$$

dove:

- y è la funzione di attivazione, che trasforma la somma ponderata degli input (più il bias) in un valore di output che rispetti determinate caratteristiche desiderate (ad esempio, limitare l'output a un intervallo specifico).
- b è detto "bias", che permette di traslare la funzione di attivazione orizzontalmente, dando al neurone la possibilità di attivarsi anche quando tutti gli input sono zero o bassi;
- w è il peso;
- x è l'input.

L'allenamento di una rete neurale permette di configurare correttamente i pesi al suo interno (w_1 , w_2 , b) attraverso algoritmi come la *backpropagation*, che permettono alla rete di adattarsi ai dati. La rete neurale, quindi, durante la fase di allenamento, impara i migliori pesi possibili per risolvere il problema che è chiamata a svolgere. Esistono due tipologie di allenamento di una rete neurale:

- *Supervised Learning*: la rete viene addestrata con dati di input associati alle rispettive etichette o soluzioni attese. La rete elabora i dati in input e confronta l'output con le soluzioni attese, correggendo i pesi se necessario. Se questi sono giusti (ovvero concordi con le soluzioni fornite), non modifica i pesi, in caso contrario, li aggiorna per ridurre l'errore.
- *Unsupervised Learning*: la rete individua autonomamente pattern o strutture latenti nei dati, senza disporre di etichette esplicite.

La struttura nella figura 2.4 è anche chiamata perceptrone, ed è la rete neurale più semplice. Combinando più perceptroni si ottiene una "Feed Forward Network".

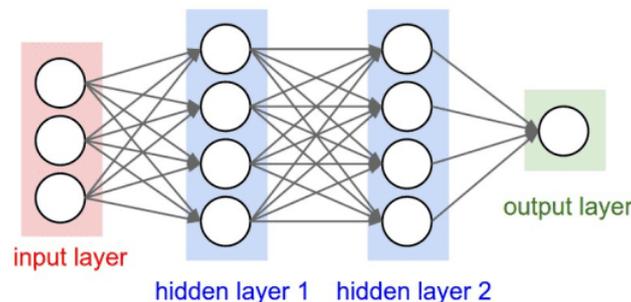


Figura 2.5: Feed Forward Network.

In questa tipologia di rete neurale, tra nodi di input e nodi di output, vi sono una serie di nodi intermedi, detti "Hidden Layers". In particolare, se una rete neurale ha più di un hidden layer, allora viene definita "Deep".

Per far avere in input delle lettere e non dei numeri, esistono due sistemi:

- *Encoding*: si ha un vettore di V zeri, dove V rappresenta la dimensione del vocabolario; ogni parola è rappresentata da un 1 al posto di uno 0, e la posizione è univoca. Questo sistema si chiama “One-hot Encoding”, ma non è una tecnica ottimale, in quanto non cattura le relazioni semantiche tra le parole.
- *Embeddings*: a ciascuna parola del vocabolario viene associato un vettore numerico, che se rappresentato nel piano cartesiano permette di ottenere il concetto di “somiglianza semantica” tra le parole. Per far ciò bisogna stabilire la dimensione dei vettori associati ad ogni parola e i valori specifici da assegnare ad ogni componente del vettore. Un metodo comune per determinare questi valori è basato sulla “Distributional Hypothesis” (o “dimmi con chi vai e ti dirò chi sei”). L’idea è che parole che appaiono in contesti simili hanno probabilmente significati simili. Questa idea viene implementata attraverso matrici di co-occorrenza, dove si contano quante volte le parole appaiono vicine nel testo.

Gli embeddings introducono due principali problematiche:

- Le matrici sono spesso “sparse” (piene di zeri), quindi si usano tecniche di riduzione dimensionale come la PCA (Principal Component Analysis) o, in alcuni casi, LDA (Linear Discriminant Analysis), per mantenere le componenti informative.
- Possono riflettere i bias presenti nei testi utilizzati per l’addestramento, quindi le risposte di queste reti neurali saranno concordi rispetto al testo usato per l’addestramento.

Nel 2013, Mikolov pensa di creare una rete neurale che crei gli embeddings, per poi sfruttarli in altre reti. Crea, quindi, una rete neurale che riceve in input vettori “one-hot” di dimensione V (detti anche “1-to- V ”) di cui si è precedentemente parlato, e restituisce in output degli embeddings molto più raffinati di quelli che si possono creare con matrici di co-occorrenza. Questi embeddings prendono il nome di “Word2vec”²².

La prima rete neurale ha in input vettori “one-hot” di dimensione V che attraversano alcuni “hidden layers” ed infine un layer “softmax” che produce un vettore di output con V valori, in cui ogni valore rappresenta la probabilità che una specifica parola sia correlata alla parola di input. L’allenamento avviene prendendo testi esistenti e, per ogni parola nel testo, si cerca di prevedere le parole vicine. Siccome la risposta è poi contenuta nel testo stesso, questa rete neurale “impara da sola”. In questo contesto, gli embeddings sono appresi come parte dei pesi all’interno degli “hidden layers”, e sono questi che, modificati durante la fase di allenamento, catturano le relazioni semantiche tra le parole. Dai pesi di questi “hidden layers” si creano gli embeddings associati, che saranno vettori di N dimensioni dove la distanza tra le parole rappresenta l’effettiva somiglianza semantica tra queste.

Considerando una dimensione degli embeddings di circa 300 componenti (valore empiricamente validato come appropriato per numerose applicazioni), lo spazio vettoriale risultante sarà caratterizzato da trecento dimensioni, rendendo impossibile una rappresentazione grafica diretta. Tuttavia, è possibile applicare tecniche di riduzione dimensionale per comprimere le distanze e proiettare lo spazio a trecento dimensioni su un piano bidimensionale: tale processo ha finalità esclusivamente visualizzative, poiché l’operazione di proiezione altera intrinsecamente il comportamento effettivo del sistema.

²²Tomas Mikolov, *Efficient Estimation of Word Representations in Vector Space*, 2013. [21]

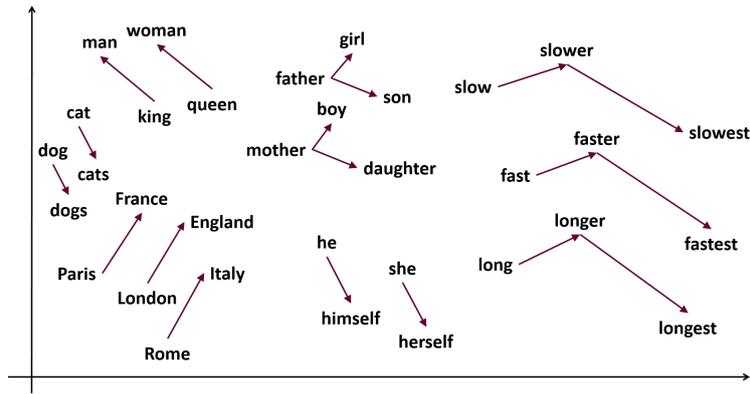


Figura 2.6: Rappresentazione bidimensionale di Word2Vec.

L'architettura Word2Vec dimostra la capacità di catturare relazioni semantiche complesse tra termini linguistici che non emergerebbero attraverso approcci tradizionali. È possibile interrogare il sistema con relazioni analogiche del tipo:

$$\text{Re} : \text{Uomo} = \text{Regina} : X$$

Il modello è in grado di identificare correttamente che $X = \text{Donna}$, dimostrando la sua capacità di riconoscere che la relazione semantica tra “Re” e “Uomo” è analoga a quella esistente tra “Regina” e “Donna”.

È fondamentale considerare che, essendo il sistema addestrato su contenuti testuali preesistenti, esso risulta inevitabilmente influenzato dai bias culturali e linguistici presenti nei dati di training. Tale caratteristica rende Word2Vec uno strumento particolarmente utile per l'analisi dei bias intrinseci della cultura di riferimento, poiché il modello evidenzia le associazioni semantiche implicite: ad esempio, termini come “donna” e “infermiera” risultano semanticamente più vicini rispetto alla coppia “uomo” e “infermiere”, riflettendo le associazioni di genere prevalenti nella letteratura utilizzata per l'addestramento. Questo è un fenomeno fondamentale da considerare quando si valuta l'output di una rete neurale, poiché le risposte generate possono riflettere e perpetuare stereotipi e pregiudizi culturali.

A questo punto si sfruttano le RNN (Recurrent Neural Network), che utilizzano un meccanismo di ricorrenza, cioè un loop interno che permette alla rete di “ricordare” le informazioni passate.

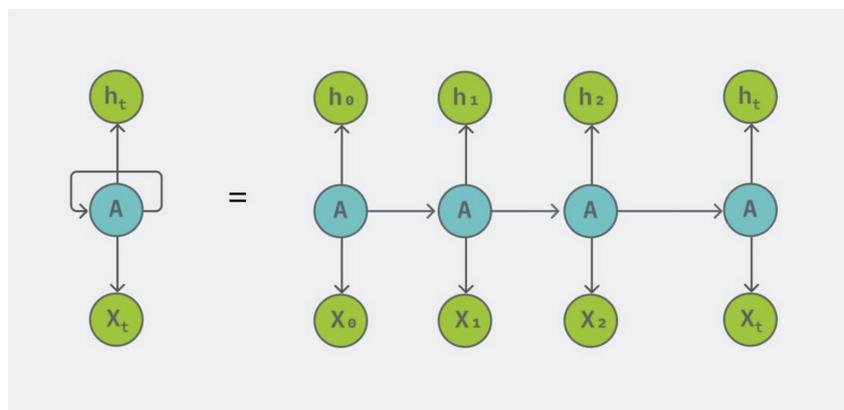


Figura 2.7: Recurrent Neural Network.

Sostanzialmente, la rete riceve in input x_t e lo stato precedente h_{t-1} e produce h_t .

$$h_t = y(w^{hh}h_{t-1} + w^{hx}x_t)$$

$$y_t = \text{softmax}(w^s h_t)$$

dove:

- h_t è la memoria attuale della RNN.
- w^{hh} e w^{hx} sono i pesi della rete.
- h_{t-1} è l'informazione della parola precedente.
- x_t è la parola attuale.
- y_t è la previsione sulla parola successiva.

Le RNN sono poi migliorate dalle LSTM (Long Short Term Memory)²³, che permettono di regolare il concetto di memoria con più precisione, permettendo la decisione di cosa ricordarsi e cosa no.

Queste LSTM sono usate all'interno delle reti Seq2Seq (Sequence-To-Sequence)²⁴, che sono un tipo di rete neurale usata per trasformare una sequenza in un'altra. Sia l'encoder che il decoder sono spesso costruiti con LSTM, un tipo di rete neurale "ricorrente" con una memoria potenziata. Le LSTM sono brave a ricordare cosa è stato detto qualche parola prima, quindi risultano ottime per lavorare con le frasi.

Talvolta, però, quando la frase è lunga o complessa, il decoder non riesce a ricordarsi bene tutto quello che ha detto l'encoder. Viene, quindi, introdotto il concetto di "Attention"²⁵, che permette al decoder di "guardare indietro" alla frase originale e decidere a quali parole prestare più attenzione in ogni momento.

Dalle LSTM si è poi passati ai *Transformers*, che mantengono l'idea dell'encoder e del decoder, ma eliminano completamente le reti neurali ricorrenti. Usano solo il meccanismo di Attention, potenziandolo con la *self-attention*, che permette a ogni parola di analizzare anche il contesto di tutte le altre parole nella frase. Questo rende la comprensione molto più profonda: se una parola ha più significati, il Transformer capisce quale intendere grazie al contesto.

Dal momento che i Transformers non hanno memoria "interna" come le LSTM, per non perdere l'ordine delle parole si usa un sistema chiamato *positional encoding*, che inserisce delle informazioni su "dove si trova" ogni parola nella frase. Così si mantiene l'ordine corretto anche senza ricorrenza.

Gli attuali LLM sono costituiti da più Transformers impilati e adottano un'architettura decoder-only²⁶, poiché il loro obiettivo principale non è comprendere un input da tradurre,

²³Sepp Hochreiter, *Long Short-Term Memory*, 1997. [22]

²⁴Ilya Sutskever, *Sequence to Sequence Learning with Neural Networks*, 2014. [23]

²⁵Ashish Vaswani, *Attention Is All You Need*, 2017. [24]

²⁶Tom B. Brown, *Language Models are Few-Shot Learners*, 2020. [25]

ma generare una continuazione coerente del testo a partire da un prompt: non ha quindi bisogno di un encoder.

2.2.2 Comparazione IA per la Generazione Testuale

Al fine di confrontare le intelligenze artificiali di text generation si farà riferimento ad un benchmark²⁷ effettuato su 32 LLM differenti, volto a identificare quale fra questi fosse il più indicato a compiere un task specifico, ovvero quello della scrittura creativa.

In particolare, questo benchmark è stato sviluppato per valutare la capacità dei Large Language Models (LLM) di integrare una serie di dieci elementi narrativi fondamentali nella produzione di testi narrativi brevi. Questa valutazione risulta particolarmente rilevante nell'ambito dell'impiego creativo degli LLM, in quanto consente di confrontare la coesione delle storie.

Questi dieci elementi narrativi sopra citati sono:

- Personaggio.
- Obiettivo.
- Concept.
- Attributi.
- Azioni.
- Metodi.
- Ambientazione.
- Tempo della storia.
- Motivazione.
- Tono.

L'adozione di un insieme eterogeneo di questi elementi, assegnati in maniera casuale, assicura che i modelli debbano generare narrazioni diversificate, riducendo il rischio di ripetizione e favorendo la variabilità dei contenuti prodotti. Il benchmark esamina sia il grado di aderenza ai vincoli imposti — verificando se il modello ha incorporato correttamente tutti gli elementi richiesti — sia la qualità letteraria del testo finale, valutandone la coerenza, la struttura e il livello di coinvolgimento.

La valutazione è stata effettuata da sei LLMs (GPT-4o, Claude 3.5 Sonnet, LLama 3.1 405B, DeepSeek-V3, Grok 2, Gemini 1.5 Pro), che avevano lo scopo di attribuire un punteggio alle singole storie sulla base di diversi parametri, ovvero:

- Sviluppo e motivazione dei personaggi.
- Struttura della trama e coerenza.
- Costruzione del mondo e atmosfera.

²⁷Lech Mazur, *LLM Creative Story-Writing Benchmark*, 2025. [26]

- Impatto narrativo e qualità della scrittura.
- Autenticità e originalità.
- Esecuzione e coesione.

Era inoltre valutata l'attinenza ai dieci elementi obbligatori, arrivando quindi ad un totale di 16 criteri di valutazione.

Ciascuno dei 32 LLMs presi in esame ha prodotto 500 brevi storie (da 400/500 parole ciascuna).

Di conseguenza, il totale dei voti generati corrisponde a:

$$32 \times 500 \times 6 \times 16 = 1,536,000$$

dove:

- “32” è il numero di LLMs testati.
- “500” è il numero brevi storie create da ciascun LLM.
- “6” è il numero di LLMs utilizzati per la valutazione.
- “16” è il numero di criteri di valutazione totali.

Di seguito è riportata una tabella con i dieci migliori LLMs in termini di punteggio medio in questo specifico task. Il punteggio è compreso tra 0 ed 1 in quanto è il range in cui veniva assegnato il punteggio dagli LLMs valutatori.

Rank	LLM	Punteggio Medio
1	DeepSeek R1	0.991
2	Claude 3.7 Sonnet Thinking 16K	0.970
3	Claude 3.5 Sonnet 2024-10-22	0.942
4	Claude 3.7 Sonnet	0.843
5	Claude 3.5 Haiku	0.541
6	Gemini 2.0 Pro Exp 02-05	0.529
7	Gemini 1.5 Pro (Sept)	0.443
8	GPT-4o Feb 2025	0.407
9	Gemini 2.0 Flash Thinking Exp Old	0.349
10	GPT-4o 2024-11-20	0.302

Tabella 2.1: Classifica delle performance dei modelli di linguaggio (LLM) secondo il benchmark

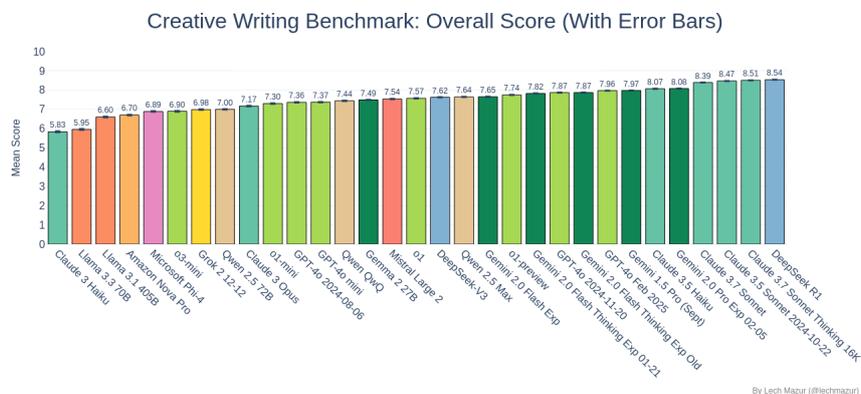


Figura 2.8: Risultati del Benchmark.

La figura 2.8 mostra i punteggi normalizzati in base dieci, aggiungendo anche una barra di errore per quella che è l'incertezza nella misurazione del punteggio.

Infine, viene mostrata una heatmap che analizza i singoli punteggi di ogni LLM per ognuno dei sedici criteri di valutazione:

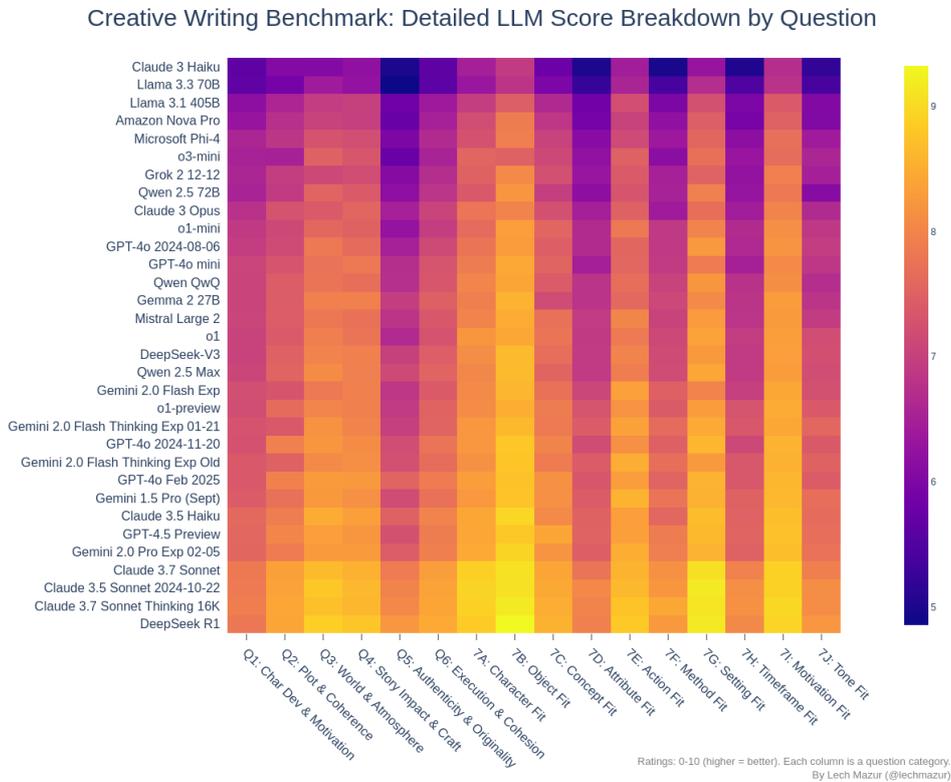


Figura 2.9: Heatmap dei singoli criteri di valutazione.

Alla luce del benchmark visionato, gli LLM più performanti (e che sono utilizzabili gratuitamente) risultano essere:

- DeepSeek R1.
- Claude 3.7 Sonnet.
- ChatGPT-4o.

2.3 IA per la Generazione di Immagini

La generazione automatica di immagini costituisce oggi una delle aree più affascinanti dell'intelligenza artificiale, grazie a un'evoluzione tecnologica che ha profondamente ampliato il potenziale creativo delle macchine. Negli ultimi decenni, questo settore ha conosciuto uno sviluppo notevole, passando da semplici algoritmi di sintesi a modelli avanzati in grado di produrre immagini di qualità fotografica, talvolta difficili da distinguere da quelle reali.

A differenza della generazione testuale, dove la valutazione può basarsi su metriche linguistiche consolidate, la generazione di immagini presenta sfide valutative uniche che richiedono criteri multidimensionali: qualità visiva, coerenza semantica, diversità creativa e fedeltà alle specifiche richieste.

Il presente paragrafo, dopo aver approfondito il funzionamento e l'evoluzione di questi sistemi, si concentra sull'analisi di un benchmark specificamente progettato per confrontare i diversi sistemi di *image generation* tramite diversi parametri, quali la qualità delle immagini prodotte, il tempo di generazione medio e il prezzo per la creazione dell'immagine. Questo benchmark ha confrontato cinquantatré modelli di generazione di immagini, fornendo una panoramica dettagliata delle prestazioni e delle capacità di ciascun sistema, e permettendo l'identificazione dei migliori sistemi in questo momento storico.

2.3.1 Funzionamento delle IA per la Generazione di Immagini

L'evoluzione dei modelli generativi per immagini può essere concettualizzata come una successione di paradigmi tecnologici, ciascuno caratterizzato da approcci architetturali distinti e da specifiche limitazioni che hanno guidato lo sviluppo delle generazioni successive. Questa progressione ha portato alla nascita di sistemi in grado di interpretare descrizioni testuali complesse e tradurle in rappresentazioni visuali coerenti, aprendo scenari applicativi che spaziano dalla produzione artistica alla sintesi di dati per l'addestramento di altri modelli di intelligenza artificiale.

Variational Autoencoders (VAE)

I Variational Autoencoders (VAE), introdotti da Kingma e Welling del 2013²⁸, hanno rappresentato una svolta nel campo dell'intelligenza artificiale generativa, in quanto rendevano possibile un modello che potesse sia comprendere la struttura dei dati che generare nuovi contenuti.

I VAE si possono vedere come un sistema composto da due parti:

- *Encoder* (codificatore), che funziona come un “compressore intelligente”: prende un'immagine e la trasforma in una rappresentazione più compatta, chiamata *spazio latente*.
- *Decoder* (decodificatore), che agisce da “decompressore”: prende la rappresentazione complessa e ricostruisce l'immagine originale.

²⁸Diederik P. Kingma and Max Welling, *Auto-Encoding Variational Bayes*, 2013. [27]

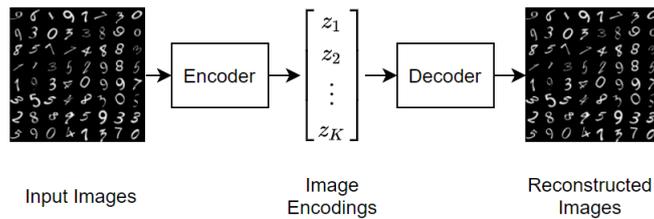


Figura 2.10: Funzionamento dei VAE.

La grande innovazione che portano i VAE è che l’encoder non produce dei numeri fissi, bensì una distribuzione probabilistica: si è passati dal concepire un’immagine come un insieme di numeri al pensarla con una propria distribuzione media e varianza standard. Questo approccio comporta due principali vantaggi:

- Gestisce l’incertezza nella rappresentazione.
- Genera nuove immagini campionando da queste distribuzioni.

L’obiettivo dei VAE è quello di rispondere alla domanda “Qual è la probabilità che questi dati siano generati dal modello?”, ma siccome calcolare questa probabilità è matematicamente intrattabile, il VAE introduce un limite inferiore (lower bound), che garantisce che la probabilità sia almeno questo valore.

La formula matematica per il calcolo del limite inferiore *ELBO* (Evidence Lower Bound) è la seguente:

$$ELBO = E_q[\log p(x|z)] - D_{KL}(q(z|x)||p(z))$$

dove:

- $E_q[\log p(x|z)]$ è il termine di ricostruzione, che misura quanto bene il decoder riesce a ricostruire l’immagine originale x . Se la ricostruzione è fedele, il valore è alto, e viceversa.
- $D_{KL}(q(z|x)||p(z))$ è il termine di regolarizzazione, che misura quanto la distribuzione appresa dall’encoder si discosta da una distribuzione “ideale” predefinita (tipicamente una gaussiana standard). Senza questo termine, l’encoder potrebbe apprendere rappresentazioni molto diverse per immagini simili. Inoltre, forza lo spazio latente ad avere una struttura ordinata e continua, permettendo di generare nuove immagini campionando punti casuali nello spazio latente.

Tuttavia, i VAE presentano alcune limitazioni significative, tra cui un problema di sfocatura. In particolare, le immagini generate dai VAE tendono ad essere sfocate per due principali motivi:

- Metriche pixel-wise: il termine di ricostruzione confronta ogni pixel individualmente usando metriche come l’errore quadratico medio.
- Eddetto di “mediazione”: quando il modello è incerto tra diverse possibilità, la funzione di loss lo penalizza meno se sceglie un valore intermedio (grigio). Questo porta a immagini “mediate” che appaiono sfocate.

Generative Adversarial Networks (GAN)

Le Generative Adversarial Networks, proposte da Goodfellow e colleghi nel 2014²⁹, introducono un paradigma completamente nuovo basato su un processo *adversarial* (competitivo) in cui due modelli neurali, un *generator* e un *discriminator*, vengono addestrati simultaneamente in una dinamica competitiva. Questa architettura segna una svolta fondamentale nella generazione di immagini, superando molte delle limitazioni intrinseche ai VAE.

In particolare:

- Il Generator (generatore) prende in input un vettore di rumore causale e lo trasforma in un'immagine sintetica realistica.
- Il Discriminator (discriminatore) riceve sia immagini reali (dal dataset di addestramento) che immagini sintetiche (generate dal generatore) e deve “giudicare” quali sono le immagini reali e quali quelle sintetiche.

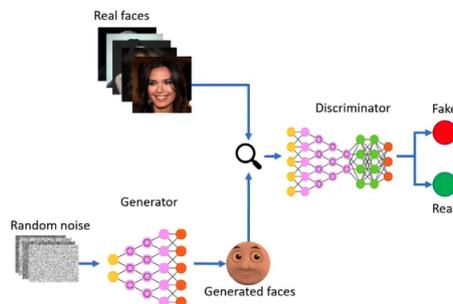


Figura 2.11: Funzionamento dei GAN.

Questo processo fa sì che, se il discriminatore trova che le immagini prodotte dal generatore sono false, i pesi del generatore vengono aggiornati. Invece, se il discriminatore sbaglia a classificare un'immagine, vengono aggiornati i pesi al suo interno.

La formula matematica che descrive questa dinamica competitiva è la seguente:

$$\min_G \max_D V(D, G) = E_x[\log D(x)] + E_z[\log(1 - D(G(z)))]$$

dove:

- $E_x[\log D(x)]$ rappresenta la probabilità che il discriminatore classifichi correttamente un'immagine reale come reale.
- $E_z[\log(1 - D(G(z)))]$ rappresenta la probabilità che il discriminatore classifichi un'immagine generata come falsa.
- \min_G indica che il generatore cerca di minimizzare questa funzione, cercando di ingannare il discriminatore.
- \max_D indica che il discriminatore cerca di massimizzare questa funzione, cercando di distinguere correttamente tra immagini reali e generate.

²⁹Ian J. Goodfellow et al., *Generative Adversarial Networks*, 2014. [28]

Questa formulazione competitiva presenta vantaggi sostanziali rispetto ai VAE. Non confrontando più le immagini “pixel per pixel” (come invece accadeva con i VAE), le GAN possono produrre immagini significativamente più nitide e realistiche. Il discriminator agisce come un critico “intelligente” che valuta la qualità globale dell’immagine piuttosto che confronti puntuali, permettendo al generator di sviluppare capacità creative più sofisticate.

Tuttavia, anche le GAN presentano alcune problematiche, tra cui:

- *Mode Collapse*: il generatore può imparare a produrre solo un numero limitato di immagini, perdendo la diversità.
- Instabilità nell’addestramento: la convergenza del processo competitivo tra le due reti neurali non è garantita, e si necessita quindi di tecniche di stabilizzazione e bilanciamento accurato tra le capacità delle due reti.

Evoluzioni delle GAN

L’architettura originale delle GAN ha subito numerose evoluzioni tecniche volte a superare le limitazioni iniziali e a migliorare sia la stabilità dell’addestramento che la qualità delle immagini generate.

- Le *Deep Convolutional GAN (DCGAN)* rappresentano una delle prime stabilizzazioni architetturali significative, introducendo linee guida specifiche per l’implementazione di reti convoluzionali nel contesto *adversarial*. Queste includono l’eliminazione dei layer completamente connessi (*fully-connected*) in favore di convoluzioni globali. Inoltre, sono utilizzate tecniche di normalizzazione per evitare che i valori diventino troppo grandi o piccoli.
- Le *Progressive GAN* introducono un nuovo approccio: invece di cercare di generare subito immagini ad alta risoluzione, la rete genera immagini piccole e sfocate e gradualmente aggiunge layer per aumentare la definizione. Questa strategia permette una stabilizzazione dell’addestramento e la produzione di immagini ad alta risoluzione precedentemente inaccessibili.
- Le *StyleGAN* introducono la possibilità di un controllo granulare sullo stile e sulle caratteristiche delle immagini generate. Questo avviene tramite l’implementazione di un *mapping network*, che trasforma il rumore latente in uno spazio di stile intermedio. Questa tipologia di GAN permette quindi la manipolazione controllata di attributi specifici dell’immagine, dalla struttura generale ai dettagli fini.
- Le *Vision Transformer GAN (ViTGAN)*³⁰ rappresentano l’integrazione più recente con le architetture Transformer, sfruttando i meccanismi di *attention* per catturare dipendenze spaziali a lungo raggio nelle immagini.

Diffusion Models

Nel 2020, Jonathan Ho, Ajay Jain e Pieter Abbeel pubblicano un paper intitolato “*Denoising Diffusion Probabilistic Models*”³¹, che introduce il concetto di *Diffusion Models* e

³⁰Kwonjoon Lee et al., *ViTGAN: Training GANs with Vision Transformers*, 2024. [29]

³¹Jonathan Ho, Ajay Jain and Pieter Abbeel, *Denoising Diffusion Probabilistic Models*, 2020. [30]

che rappresenta probabilmente l'evoluzione più significativa nel campo della generazione di immagini.

Questi modelli di diffusione sono ispirati alla fisica statistica e si basano su due processi complementari: un processo di diffusione diretto che aggiunge progressivamente rumore gaussiano ai dati reali fino a renderli indistinguibili da rumore puro, e un processo di diffusione inverso che apprende a rimuovere gradualmente questo rumore per ricostruire immagini coerenti.

In particolare:

- Il processo di diffusione diretto è un processo markoviano (processo in cui per prevedere il successivo risultato è sufficiente conoscere il risultato presente) che trasforma gradualmente un'immagine reale x_0 in rumore gaussiano attraverso T step temporali. A ogni step t , viene aggiunto rumore gaussiano secondo una schedulazione predefinita, creando una sequenza x_1, x_2, \dots, x_T dove x_T rappresenta l'immagine progressivamente più rumorosa.
- Il processo di diffusione inverso costituisce il cuore dell'apprendimento del modello: la rete neurale viene addestrata a predire e rimuovere il rumore aggiunto a ogni step temporale, apprendendo essenzialmente a "invertire" il processo di corruzione. Questo addestramento avviene attraverso un obiettivo di *denoising score matching*, dove il modello apprende a stimare il gradiente della distribuzione di probabilità dei dati rumorosi.

L'implementazione tecnica dei modelli di diffusione richiede la definizione precisa dei processi diretti e inversi attraverso catene markoviane parametrizzate.

Il processo diretto è definito come:

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1})$$

dove ogni transizione $q(x_t|x_{t-1})$ rappresenta l'aggiunta di rumore gaussiano con varianza crescente secondo una schedulazione predefinita.

Il processo inverso approssimato è parametrizzato come:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t)$$

dove p_θ rappresenta la rete neurale che apprende a predire la distribuzione condizionale inversa.

L'architettura neurale più comunemente utilizzata è basata su *U-Net*, quindi una rete neurale in cui:

- Una parte (Encoder) analizza l'immagine riducendola e catturando le caratteristiche principali.
- Una parte (Decoder) ricostruisce l'immagine ingrandendola e aggiungendo dettagli.
- Le due parti sono collegate da connessioni *skip* che permettono di mantenere informazioni dettagliate durante la ricostruzione.

Questa tipologia di architettura è particolarmente efficace, in quanto può lavorare su diversi livelli di dettaglio contemporaneamente, non perde informazioni importanti ed è specializzata (in quanto nata per rimuovere rumore dalle immagini).

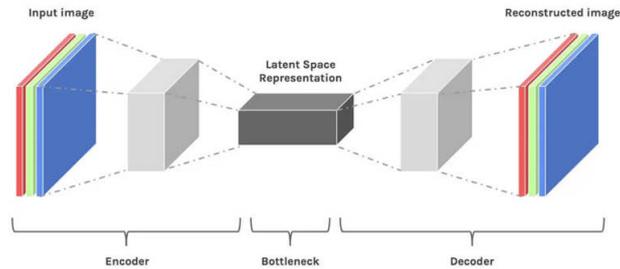


Figura 2.12: Funzionamento dei modelli di diffusione.

Siccome il modello deve conoscere anche a quale step temporale si trova, vengono incorporati degli *embedding posizionali*, che codificano lo step temporale corrente, permettendo al modello di adattare la sua strategia di denoising in base al livello di rumore presente.

Inoltre, sono integrati nell'architettura U-Net anche meccanismi di *attention*, al fine di catturare dipendenze spaziali a lungo raggio e permettere al modello di mantenere coerenza globale durante tutto il processo di denoising.



Figura 2.13: Da sinistra a destra: GAN, Stable Diffusion, Immagini per l'addestramento.

I modelli di diffusione offrono diversi vantaggi rispetto alle architetture generative precedenti:

1. Stabilità nell'addestramento: come visto in precedenza, questa rappresentava una problematica significativa nel caso delle GAN; in questo modello non vi è più competizione tra due reti, ma solo un processo di denoising.
2. Qualità più elevata delle immagini³²: il processo graduale di generazione e il controllo sistematico del rumore fanno sì che le immagini generate dai modelli di diffusione abbiano una maggiore risoluzione, coerenza spaziale e fedeltà ai dettagli, come mostrato nella figura 2.13. In particolare risultano particolarmente efficienti nella generazione di texture complesse, illuminazione realistica e coerenza strutturale globale.
3. Diversità nelle generazioni: i modelli di diffusione mostrano una maggiore varietà nelle immagini generate, evitando il problema del *mode collapse* che invece affligge diverse

³²Prafulla Dhariwal e Alex Nichol, *Diffusion Models Beat GANs on Image Synthesis*, 2021. [31]

architetture GAN. Questo è dovuto alla natura stocastica del processo di diffusione, che permette una più ampia gamma di possibili output.

4. Controllo avanzato: è possibile guidare la generazione attraverso testi descrittivi, immagini di riferimento o combinazioni di condizioni multiple, permettendo quindi un controllo creativo che non era precedentemente accessibile.

2.3.2 Confronto IA per la Generazione di Immagini

Al fine di confrontare le diverse IA di image generation si farà riferimento all'analisi dettagliata effettuata dal portale ArtificialAnalysis³³, che ha confrontato cinquantatré diversi modelli di generazione di immagini sulla base di tre criteri principali:

- Qualità.
- Tempo di generazione.
- Prezzo.

In quest'analisi mancherà, però, una valutazione di Midjourney, in quanto questo modello non offre un'API pubblica per l'accesso diretto, ma è utilizzabile solamente tramite un altro applicativo, ovvero Discord. Sebbene sia assente nelle valutazioni, bisogna tener conto della rilevanza nel settore, in quanto è uno dei modelli più noti in ambito artistico e creativo.

Si procede ora ad approfondire ciascuno dei tre criteri, riportando, per ognuno, i modelli che hanno ottenuto i risultati migliori.

Qualità

La qualità delle immagini generate è quantificata attraverso un punteggio definito ELO, assegnato a ciascun modello sulla base delle preferenze espresse dagli utenti nella cosiddetta "Image Arena", un ambiente virtuale nel quale diversi modelli competono producendo immagini a partire dallo stesso prompt testuale. La valutazione avviene tramite test comparativi che hanno coinvolto oltre 100.000 risposte di utenti reali, riflettendo quindi in modo affidabile la capacità del modello di produrre risultati realistici e fedeli alla descrizione fornita. Di seguito è riportato un grafico del punteggio ELO attribuito ai diversi modelli:

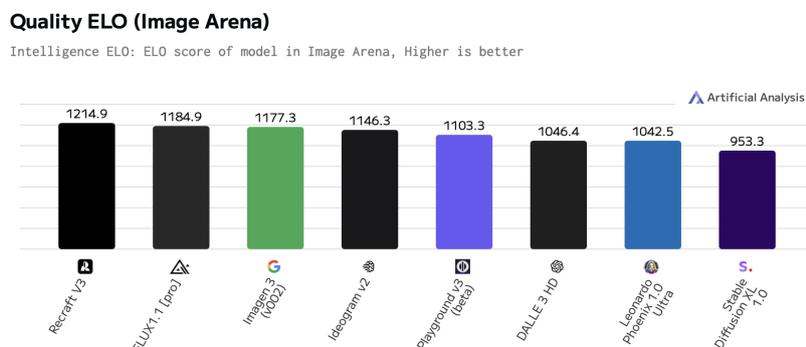


Figura 2.14: Confronto sul punteggio ELO.

³³Artificial Analysis, *Text-to-Image Models Leaderboard*, 2025. [32]

Questi dati trovano riscontro anche nella percentuale di vittorie nella “Image Arena”, i cui dati sono i seguenti:

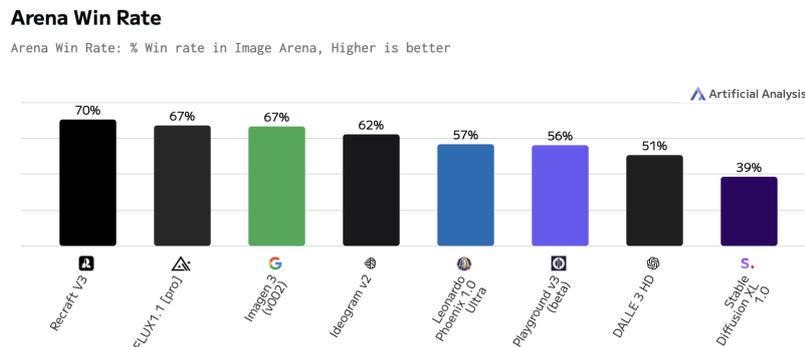


Figura 2.15: Confronto sulla percentuale di vittoria nell’“Image Arena”.

Dai grafici emerge come *Recraft V3*, *FLUX1.1 (Pro)* e *Imagen 3* siano i modelli che hanno ottenuto i punteggi migliori in termini di qualità, sia come valutazione ELO che come percentuale di vittorie nella “Image Arena”.

Tempo di Generazione

Questo criterio misura il tempo medio necessario al modello per produrre una singola immagine a partire da un input testuale.

Un’analisi effettuata durante le ultime due settimane riporta questi risultati:

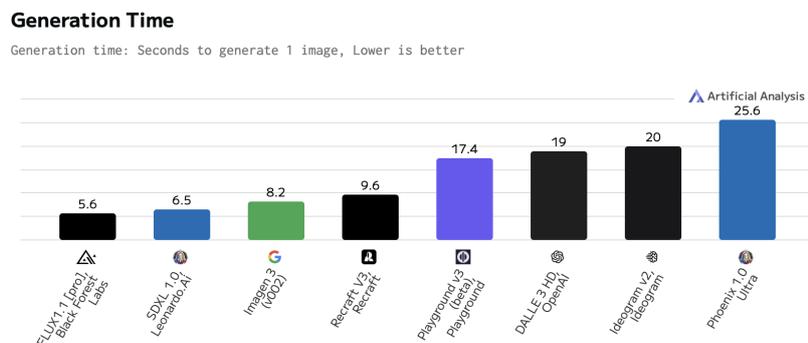


Figura 2.16: Confronto sul tempo di generazione medio per un’immagine.

Tenendo conto della varianza, ovvero i valori di quanto i tempi possano variare rispetto alla media, si può analizzare la stabilità di questi modelli. Di seguito è riportato il grafico che la analizza:

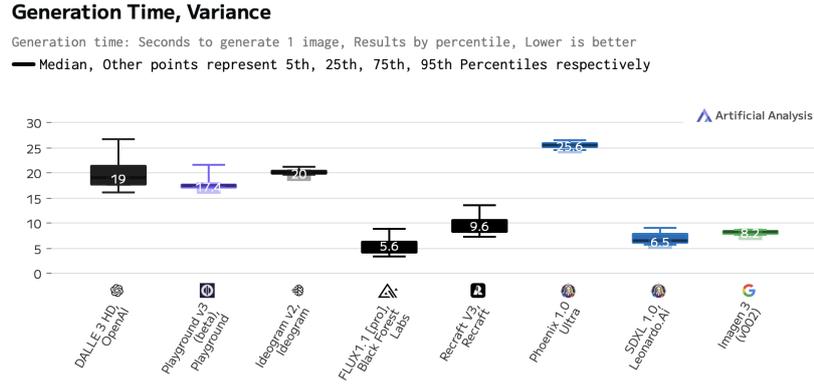


Figura 2.17: Confronto sul tempo di generazione comprensivo di varianza per un'immagine.

Partendo da questi dati risulta che *Flux1.1 (Pro)* sia mediamente il più veloce, seguito da *LeonardoAI*, *Imagen 3* e *Recraft V3*. Il più stabile (in termini di varianza) risulta essere *Imagen 3*.

Prezzo

Il prezzo misura il costo effettivo in dollari per la generazione di mille immagini, ed è un fattore cruciale in produzioni in cui il budget non risulti elevato.

Di seguito riporto il grafico con i prezzi dei vari modelli:

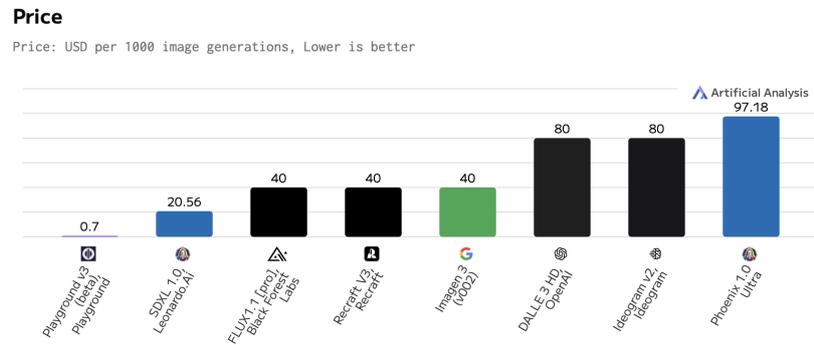


Figura 2.18: Confronto sul prezzo in dollari per la generazione di mille immagini.

I modelli che risultano costare meno per effettuare mille generazioni risultano essere:

- *Playground v3*, con un costo di 0.7 dollari.
- *LeonardoAI*, con un costo di 20.56 dollari.
- *Flux1.1 (Pro)*, con un costo di 40 dollari.
- *Recraft V3*, con un costo di 40 dollari.
- *Imagen 3*, con un costo di 40 dollari.

2.4 IA per la Generazione di Video

La *video generation* rappresenta oggi l'evoluzione più complessa dell'intelligenza artificiale generativa, rappresentando una sintesi delle sfide tecnologiche affrontate nei casi di generazione testuale e di immagini. La caratteristica principale che la generazione video introduce è la dimensione temporale: non si deve più creare un'immagine statica, che vive in uno spazio a due dimensioni (larghezza \times altezza), ma questa deve variare nel tempo, aggiungendo quindi la terza dimensione (larghezza \times altezza \times tempo). Questa variazione amplifica esponenzialmente la complessità computazionale e concettuale.

Il presente paragrafo si propone di esplorare il funzionamento dei modelli di generazione video, partendo da un'analisi dell'evoluzione di quelli legati alla generazione di immagini e di come questi possano includere la dimensione temporale al loro interno. In seguito, verranno analizzate le problematiche principali legate a questa tecnologia, si proporrà un framework applicativo per l'utilizzo dei modelli in locale, ed infine, si andranno a confrontare le principali tecnologie di video generation attualmente presenti sul mercato, al fine di fornire una panoramica dettagliata dello stato dell'arte di questa tecnologia.

2.4.1 Funzionamento delle IA per la Generazione di Video

Come analizzato nel precedente capitolo, i modelli di intelligenza artificiale generativa basati sul deep learning (VAE, GAN, Diffusion Models), hanno mostrato risultati notevoli nella generazione di immagini statiche realistiche.

Tuttavia, l'applicazione dell'IA generativa alla creazione video presenta sfide uniche e complesse. Le proprietà spazio-temporali intrinseche dei video, la necessità di generare scene dinamiche fotorealistiche e i considerevoli costi computazionali associati all'elaborazione di dati video rappresentano ostacoli significativi.

I modelli di intelligenza artificiale generativa hanno adottato diverse strategie³⁴ per affrontare queste sfide, in particolare:

- Tra il 2018 e il 2019, le VAEs (Variational Autoencoders) sono state adattate per gestire la struttura temporale dei video tramite un condizionamento dei frame successivi a partire dai frame precedenti. Questo, ad esempio, è il concetto alla base di SVG (Stochastic Video Generation), uno dei modelli che adatta le VAEs per la *video generation*, e che introduce una gerarchia di variabili latenti, al fine di poter catturare la natura multiscala dei dati video e poter quindi generare sequenze diverse e realistiche, con una fusione fluida dei fotogrammi.

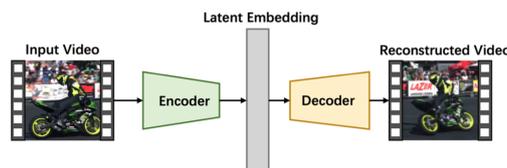


Figura 2.19: Funzionamento dei VAE applicati alla generazione video.

³⁴Pengyuan Zhou et al., *A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming*, 2024. [33]

- Le GANs (Generative Adversarial Networks) sono state modificate al fine di ottenere consistenza temporale e generare video realistici a partire dal 2016/2017. Un esempio di modello per la generazione di video che utilizza le GANs è *VideoGAN*, nel quale è introdotta una nuova architettura a due vie, per poter separare la parte di visualizzazione da quella del movimento. In questo contesto, il generatore produce i frame del video, mentre il discriminatore ne valuta il realismo ed il movimento tra due istanti consecutivi. Questo processo sfrutta quindi il carattere creativo e competitivo del modello e permette la comprensione e replica di pattern complessi, producendo video realistici.

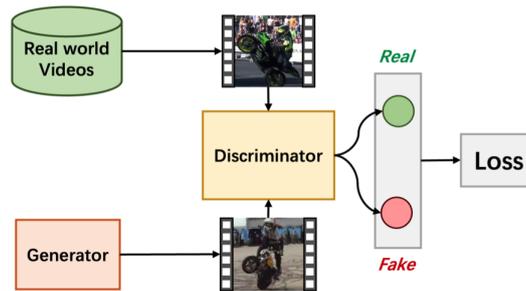


Figura 2.20: Funzionamento dei GAN applicata alla generazione video.

- Gli *Autoregressive Models* generano frame in maniera sequenziale, in modo tale che la generazione di ogni frame sia condizionata dal precedente frame. Ciò permette un'ottima continuità visiva e narrativa. Le sperimentazioni su questo tipo di modello hanno visto un'accelerazione nel periodo 2020/2021.

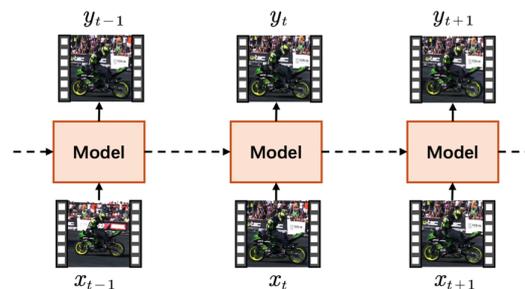


Figura 2.21: Funzionamento degli Autoregressive Models nella generazione video.

- I Diffuse Models estendono il processo di diffusione da singole immagini a sequenze temporali coerenti e rappresentano l'evoluzione più recente nel campo, in quanto risalgono al periodo tra il 2022 e il 2024. Mentre nel dominio bidimensionale il rumore gaussiano viene aggiunto e rimosso da singoli frame statici, nel contesto video il processo deve considerare la correlazione temporale tra frame consecutivi, garantendo che la progressiva rimozione del rumore mantenga coerenza narrativa e continuità visiva. Questi modelli permettono di convertire complessi prompt testuali in sequenze video ad alta risoluzione.

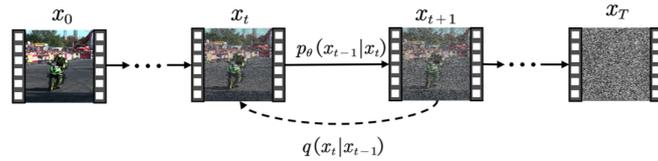


Figura 2.22: Funzionamento dei Diffuse Models nella generazione video.

Si prenderà ora in esame il caso specifico dei Diffuse Models, in quanto rappresentano l'approccio più promettente e attualmente dominante nel panorama della video generation, come dimostrato dal successo di modelli come SORA di OpenAI e Veo3, che saranno analizzati nei paragrafi successivi.

Le problematiche specifiche che emergono nel caso della generazione video per i modelli di diffusione includono:

- *Consistenza temporale*: il mantenimento della coerenza tra frame consecutivi rappresenta una sfida algoritmicamente complessa, poiché piccole variazioni nel processo di denoising possono propagarsi e amplificarsi lungo la sequenza temporale, creando artefatti visivi e discontinuità narrative.
- *Coerenza di movimento*: la rappresentazione e generazione di movimento realistico richiede la comprensione e modellazione delle dinamiche fisiche, delle traiettorie spaziali e delle relazioni causali tra oggetti in movimento.
- *Scalabilità computazionale*: l'espansione dello spazio dei parametri dal bidimensionale al tridimensionale comporta una crescita esponenziale dei requisiti computazionali, rendendo necessarie strategie di ottimizzazione e parallelizzazione specifiche.

Video Diffusion Models

I Video Diffusion Models³⁵ rappresentano l'estensione più diretta dei modelli di diffusione dal dominio delle immagini a quello video. Questa architettura mantiene i principi fondamentali dei diffusion models bidimensionali, estendendoli per gestire la dimensione temporale attraverso una reinterpretazione tridimensionale del processo di diffusione.

Il processo di diffusione viene applicato a sequenze di frame video, dove il rumore gaussiano viene aggiunto simultaneamente a tutti i frame della sequenza secondo una schedulazione temporale coerente. Durante la fase di addestramento, il modello apprende a rimuovere progressivamente questo rumore, sviluppando una comprensione intrinseca delle correlazioni spazio-temporali.

L'implementazione tecnica richiede l'utilizzo di convoluzioni tridimensionali che operano simultaneamente sulle dimensioni spaziali e temporali. Questi operatori convoluzionali tridimensionali permettono di catturare pattern locali sia nello spazio che nel tempo, modellando efficacemente fenomeni come il movimento di oggetti, i cambiamenti di illuminazione e le transizioni di scena.

³⁵Jonathan Ho et al., *Video Diffusion Models*, 2022. [34]

La formulazione matematica estende quindi il processo di diffusione bidimensionale:

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1})$$

dove ora x_0 rappresenta una sequenza video completa anziché una singola immagine, e ogni transizione $q(x_t|x_{t-1})$ preserva le correlazioni spazio-temporali attraverso kernel convoluzionali tridimensionali.

Diffusion Transformers (DiT)

I Diffusion Transformers³⁶ hanno rivoluzionato il campo della generazione video, introducendo un approccio completamente nuovo alla sintesi di contenuti visuali. Sora, il modello sviluppato da OpenAI, ne è l'esempio più emblematico: sfrutta l'architettura DiT (Diffusion Transformer) per processare simultaneamente informazioni spaziali e temporali attraverso patches estratti da video e immagini codificati nello spazio latente.

Il cuore di questa innovazione sta nel modo in cui i dati visuali vengono trattati: invece di considerare i video come sequenze continue di pixel, Sora li scompone in patches spazio-temporali che diventano i "token" del sistema, proprio come le parole in un modello di linguaggio. Questo rappresenta un cambio di paradigma rispetto alle tradizionali reti convoluzionali: ora sono i meccanismi di attention a gestire le complesse relazioni che legano spazio e tempo.

La trasformazione dei dati video in token avviene attraverso un processo articolato. Prima di tutto, ogni singolo frame viene suddiviso in patches spaziali quadrati, generalmente di 16×16 o 32×32 pixel. Questi patches vengono poi organizzati temporalmente: quelli che occupano la stessa posizione in frame consecutivi si raggruppano per formare unità spazio-temporali coerenti.

Ogni patch così ottenuto viene quindi proiettato in uno spazio vettoriale ad alta dimensionalità tramite layer lineari addestrabili, acquisendo una rappresentazione densa e ricca di informazioni. A questo punto entra in gioco l'encoding posizionale, che arricchisce ogni patch con informazioni precise sulla sua collocazione sia spaziale (dove si trova nel frame) che temporale (quando appare nella sequenza).

Il Transformer può così operare su questa rappresentazione strutturata, utilizzando l'auto-attenzione per catturare dipendenze anche molto distanti nel tempo e nello spazio. La flessibilità di questo approccio supera le limitazioni delle convoluzioni tridimensionali tradizionali, permettendo di gestire video di durata variabile e con risoluzioni diverse senza modifiche architetturali sostanziali.

Gestione della Coerenza Temporale

Le diverse architetture di modelli per la generazione dei video devono comunque occuparsi del problema della coerenza temporale dei frame. Per farlo, sono stati introdotti diversi meccanismi per la gestione di questa problematica:

³⁶William Peebles e Saining Xie, *Scalable Diffusion Models with Transformers*, 2022. [35]

- *Attention Cross-Frame*: Una delle principale soluzioni a questo problema sono i meccanismi di attention cross-frame, i quali implementano connessioni esplicite tra frame temporalmente adiacenti, permettendo la propagazione di informazioni contestuali attraverso la dimensione temporale. Dal punto di vista tecnico, questo si traduce nell'uso di maschere di attention che collegano patches corrispondenti attraverso frame adiacenti. Un oggetto che si muove nella scena mantiene così la sua identità visiva perché le informazioni sul suo aspetto si propagano naturalmente da un frame all'altro, creando quella continuità che l'occhio umano percepisce come movimento naturale.
- *Convolutioni Temporal*: questi operatori lavorano esclusivamente sulla dimensione del tempo, utilizzando kernel monodimensionali che scansionano la sequenza alla ricerca di pattern ricorrenti. Sono particolarmente bravi nel catturare quei ritmi sottili che caratterizzano il movimento naturale, ad esempio l'oscillazione di una foglia, l'accelerazione di un'auto, il rallentamento di una caduta. L'integrazione di questi layer nell'architettura complessiva avviene attraverso moduli dedicati che processano le rappresentazioni latenti lungo l'asse temporale. I filtri appresi durante l'addestramento diventano così dei "rilevatori" di dinamiche temporali tipiche, permettendo al modello di riprodurre movimenti che seguono le leggi fisiche del mondo reale.
- *Modellazione del Movimento e Integrazione dell'Optical Flow*: Le tecniche più avanzate integrano anche la comprensione del movimento nell'architettura generativa. Per farlo si utilizza un campo vettoriale bidimensionale che descrive il movimento apparente di pixel tra frame consecutivi, chiamato *Optical Flow*. L'integrazione dell'optical flow nelle architetture generative avviene attraverso tre componenti principali:
 1. Si utilizzano moduli neurali specializzati nella predizione di campi di optical flow, chiamati *Flow Prediction Networks*.
 2. I campi predetti in precedenza vengono poi utilizzati per guidare il processo di generazione, attraverso un metodo chiamato *Flow-Guided Sampling*, che assicura che ogni nuovo frame rispetti le leggi cinematiche stabilite da quelli precedenti.
 3. Le inconsistenze nei campi di movimento predetti sono penalizzate da funzioni di perdita, attraverso il processo del *Flow Consistency Losses*.

Il risultato di questa integrazione dell'optical flow permette quindi l'ottenimento di sequenze video che rispettino i principi cinematici fondamentali, andando perciò a migliorare la plausibilità percettiva del contenuto generato.

2.4.2 Framework Applicativo: ComfyUI

Come verrà analizzato nel paragrafo 3.1, la problematica principale della generazione video tramite intelligenza artificiale generativa (affinché questa diventi un'opzione anche per i grandi mercati) è la mancanza di un controllo granulare sulle generazioni.

Per superare questa limitazione, si stanno affermando soluzioni basate su approcci locali, ovvero sistemi che permettono di eseguire le generazioni direttamente sul proprio computer, eliminando la dipendenza da servizi cloud e intermediari esterni. Questo approccio garantisce maggiore controllo sui parametri di generazione e sui dati processati. Il sistema che ad oggi si avvicina maggiormente al restituire risultati accettabili e con un buon livello di controllo è basato proprio su questa filosofia locale.

Un framework essenziale per questa tipologia di processo è *ComfyUI*³⁷, un'interfaccia grafica che vuole rendere più accessibile e interagibile il processo di generazione tramite IA. La peculiarità di quest'applicativo è la sua logica basata su nodi e grafi, costruita sul modello di *Stable Diffusion*, che fornisce agli utenti una maggiore flessibilità e una comprensione più approfondita su ogni fase del processo rispetto alle interfacce tradizionali online.

ComfyUI è stato pensato per poter ospitare ed utilizzare in locale i maggiori modelli di diffusione per la generazione di immagini o video. Il sistema permette di costruire cicli di iterazioni e variazione, rendendo particolarmente semplice la comprensione di come funzionino i diversi componenti.

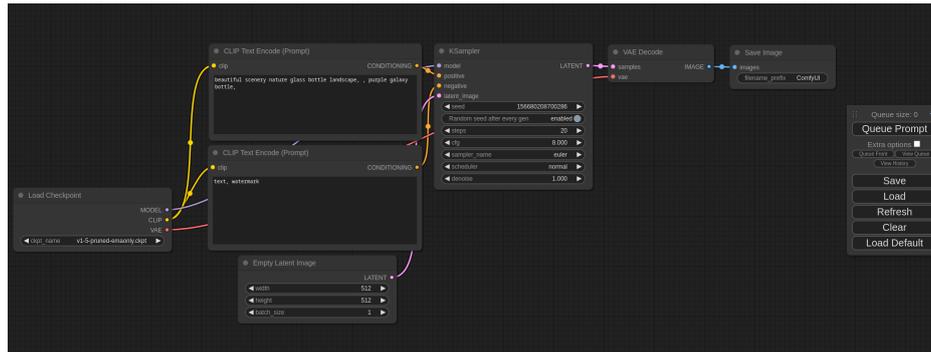


Figura 2.23: Esempio di albero di nodi su ComfyUI.

È possibile caricare il modello che si adatta meglio al task che si deve andare a compiere, rendendo quindi l'applicativo estremamente versatile.

Oltre al controllo granulare che offre, ComfyUI integra un ampio ventaglio di funzionalità avanzate che lo rendono una soluzione completa per la generazione di contenuti:

- Upscaling della qualità dei video e delle immagini.
- Generazione di immagini tramite Text-to-Image e Image-to-Image, includendo poi funzioni avanzate come l'*inpainting*, utile a modificare parti specifiche delle immagini, e il *face detailer*, utilizzato per correggere automaticamente i volti nelle immagini generate.
- Generazione di video tramite Text-to-Video e Image-to-Image Video.
- Supporta combinazioni di modelli, quindi anche l'uso di LoRA (Low-Rank Adaptation) per poter realizzare un fine-tuning efficiente dei modelli e combinare diverse influenze artistiche per personalizzare i risultati.

Il lato negativo di questa tipologia di workflow risiede nelle sue caratteristiche stesse: generando gli output in locale è necessario che la macchina che processa l'operazione sia sufficientemente potente, al fine di non incorrere in tempi di generazione lunghi o di non riuscire a portare a termine il compito. Purtroppo le caratteristiche hardware richieste per non incontrare problemi di questo tipo sono molto costose, motivo per il quale questo sistema è utilizzato per lo più per la generazione di immagini oppure in case di produzione che dispongono di un cospicuo budget.

³⁷Comfy Org, *ComfyUI*, 2023. [36]

2.4.3 Comparazione IA per la Generazione di Video

Al giorno d’oggi esistono molte IA per la generazione di video, quindi fare una comparazione che le includa tutte è molto complesso. Al fine di effettuare comunque un’analisi utile si andranno ad analizzare i principali sistemi attualmente presenti sul mercato.

Attualmente, la maggior parte di questi sistemi di generazione video utilizza dei modelli di diffusione video basati su architetture transformers.

Nello specifico si andranno ad analizzare quattro diversi tool:

- Google Veo3³⁸.
- Sora (OpenAI)³⁹.
- Runway ML⁴⁰.
- Adobe Firefly⁴¹.

Per effettuare questo confronto si utilizzeranno delle tabelle che riassumeranno le funzionalità delle varie IA. In particolare, ciascuna di queste tabelle vuole mettere in evidenza informazioni diverse:

1. Tabella 2.2: mostra una comparazione tra i modelli, mettendo in evidenza la risoluzione permessa e le varie funzioni di input e output disponibili.
2. Tabella 2.3: confronta la durata del video generato, i vari aspect-ratio disponibile e l’eventuale presenza di preset di stili visivi.
3. Tabella 2.4: prende in esame la funzione *text-to-video*, analizzandone la qualità, la disponibilità dei movimenti di camera e il frame rate disponibile.
4. Tabella 2.5: prende in esame la funzione *image-to-video*, analizzandone la qualità, la disponibilità dei movimenti di camera e il frame rate disponibile.
5. Tabella 2.6: prende in esame la funzione *image-to-image video*, analizzandone la qualità, la disponibilità dei movimenti di camera e il frame rate disponibile.

³⁸Google, *Vevo 3*, 2025. [37]

³⁹OpenAI, *Sora*, 2024. [38]

⁴⁰Runway Inc., *RunwayML*, 2025. [39]

⁴¹Adobe Inc. *Adobe Firefly*. 2025. [40]

Nome	Modello	Risoluzione	Input e Output consentiti
Google Veo3	Veo 3	Non impostabile	Text-to-Video, Image-to-Video
Sora	SORA	480p, 720p, 1080p	Text-to-Video, Image-to-Video
Runway ML	Gen-4	720p, 1080p	Text-to-Video, Image-to-Video, Image-to-Image Video
Adobe Firefly	Firefly Video Model	540p, 720p, 1080p	Text-to-Video, Image-to-Video, Image-to-Image Video

Tabella 2.2: Confronto tra modelli, risoluzione e input/output disponibili.

Nome	Durata Video	Aspect Ratio	Elenco Stili Visivi
Google Veo3	5 secondi	16:9, 1:1, 3:4, 9:16	Non presente
Sora	5, 10, 15, 20 secondi	16:9, 3:2, 1:1, 2:3, 9:16	Presente
Runway ML	5, 10 secondi	21:9, 16:9, 4:3, 1:1, 3:4, 9:16	Presente
Adobe Firefly	5 secondi	16:9, 1:1, 9:16	Presente

Tabella 2.3: Confronto tra durata dei video, aspect ratio e disponibilità preset.

Nome	Text-to-video	Qualità	Movimenti di Camera	Frame Rate
Google Veo3	Si	Ottima	Da specificare nel prompt	Non impostabile
Sora	Si	Ottima	Da specificare nel prompt	Non impostabile
Runway ML	Si	Ottima	Da specificare nel prompt	Non impostabile
Adobe Firefly	Si	Ottima	Zoom, Pan, Tilt	24 fps

Tabella 2.4: Confronto sulle funzionalità *Text-to-Video*.

Nome	Image-to-Video	Qualità	Movimenti di Camera	Frame Rate
Google Veo3	Si	Ottima	Da specificare nel prompt	Non impostabile
Sora	Si	Buona, spesso risulta imprecisa	Da specificare nel prompt	Non impostabile
Runway ML	Si	Buona	Da specificare nel prompt	Non impostabile
Adobe Firefly	Si	Buona	Zoom, Pan, Tilt	24 fps

Tabella 2.5: Confronto sulle funzionalità *Image-to-Video*.

Nome	Image-to-Image Video	Qualità	Movimenti di Camera	Frame Rate
Google Veo3	No	/	/	Non impostabile
Sora	No	/	/	Non impostabile
Runway ML	Si	Buona	Non impostabile	Non impostabile
Adobe Firefly	Si	Buona	Non impostabile	24 fps

Tabella 2.6: Confronto sulle funzionalità *Image-to-Image Video*.

Inoltre, alcune di queste intelligenze artificiali introducono funzioni proprie, come:

- Veo 3 è stato il primo modello ad introdurre la possibilità di generazione del suono in simultanea alla generazione del video, rivoluzionando quindi le prospettive future, aprendo alla possibilità di generazione di dialoghi tra persone.
- Sora ha incorporata la funzione di *storyboard*, che permette l'affiancamento di diverse “cards” contenenti ciascuna un video. Ciò permette di disporre i video all'interno di queste “cards” nella timeline e, lasciando dei vuoti tra i video, il sistema genererà connessioni visive tra le scene. Meno spazio vuoto viene lasciato, e più il taglio risulterà netto.

L'analisi comparativa condotta su questi quattro sistemi di generazione video basati su intelligenza artificiale ha evidenziato un panorama tecnologico in rapida evoluzione, caratterizzato da approcci differenti ma convergenti verso obiettivi comuni di qualità e versatilità.

Dal confronto emerge chiaramente come Google Veo3 si distingua per l'innovazione nella generazione simultanea di audio e video, rappresentando un significativo passo avanti verso la creazione di contenuti multimediali completi. Tuttavia, la mancanza di controlli granulari su parametri tecnici come risoluzione e frame rate potrebbe limitarne l'applicabilità in contesti professionali specifici.

Sora di OpenAI dimostra la maggiore flessibilità in termini di durata video (fino a 20 secondi) e varietà di aspect ratio, oltre a introdurre funzionalità innovative come lo storyboard per la creazione di narrazioni complesse. La qualità della funzione text-to-video risulta eccellente, mentre l'image-to-video mostra alcune imprecisioni che potrebbero influire sulla consistenza del risultato finale.

Runway ML si posiziona come una soluzione equilibrata, offrendo la gamma più ampia di aspect ratio e mantenendo una buona qualità su tutte le funzionalità supportate. La presenza della funzione image-to-image video lo rende particolarmente versatile per workflow creativi complessi.

Adobe Firefly si distingue per il controllo preciso sui movimenti di camera (Zoom, Pan, Tilt) e per il frame rate fisso a 24 fps, caratteristiche che lo rendono particolarmente adatto a produzioni che richiedono standard cinematografici. L'integrazione nell'ecosistema Adobe rappresenta un vantaggio significativo per i professionisti del settore.

È importante sottolineare che le valutazioni qualitative presentate in questa analisi (“Ottima”, “Buona”) sono di natura soggettiva e basate su criteri specifici relativi al progetto di tesi sviluppato, che sarà dettagliatamente illustrato nel Capitolo 3.

2.5 IA per la Generazione di Musica, Suoni e Voci

In maniera concettualmente simile al funzionamento dei modelli di intelligenza artificiale generativa per la creazione di video, anche le IA di *sound, voice e music generation* hanno come caratteristica peculiare la necessità di una coerenza interna che si sviluppi lungo l'asse del tempo. Questa ha spesso livelli multipli di rappresentazione: dalla struttura globale ai dettagli microtimbrici del suono.

In questo paragrafo si vuole esplorare il funzionamento dei diversi modelli per la sintesi di musica, suoni e voci, evidenziando l'evoluzione storica avvenuta. In seguito, saranno confrontate le principali tecnologie legate a ciascuno di questi campi, cercando quindi di identificare i migliori sistemi disponibili in questo momento storico.

2.5.1 Funzionamento dei Modelli per la Generazione Musicale

Similarmente alle tecnologie analizzate in precedenza, anche la generazione musicale tramite intelligenza artificiale generativa ha attraversato un periodo di evoluzione significativo negli ultimi anni, grazie all'introduzione di architetture basate su *deep neural networks*. Seguendo la stessa analisi già operata per i modelli di generazione video, si analizzeranno ora le diverse modalità dei modelli di IA per la generazione di musica:

- I modelli autoregressivi, come *MuseNet*⁴² di OpenAI, si basano su architetture Transformer addestrate a predire il prossimo evento musicale in una sequenza, dato il contesto dei precedenti. Questo approccio si dimostra particolarmente efficace nella gestione di strutture musicali complesse, anche su scale temporali lunghe. Tali modelli sono capaci di generare musica coerente con uno stile o genere specifico, adattandosi ad elementi come strumento, tonalità o ritmo.
- Anche l'approccio tramite *Variational Autoencoders* si è adattato per poter generare questa tipologia di contenuto, come il modello *MusicVAE*⁴³. In questo contesto, l'encoder comprime una frase musicale in uno spazio latente continuo, mentre il decoder è addestrato a ricostruire la sequenza originaria. Questo permette di esplorare lo spazio musicale in modo controllato, interpolando tra stili o creando transizioni fluide tra temi musicali. I VAEs risultano, quindi, particolarmente utili per la generazione di variazioni melodiche o per la manipolazione creativa della forma musicale.

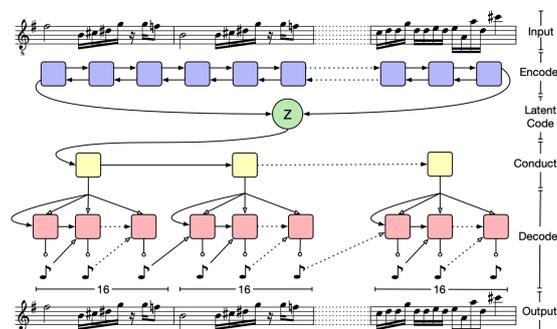


Figura 2.24: Funzionamento dei Variational Autoencoders nella generazione musicale.

⁴²OpenAI, *MuseNet*, 2019. [41]

⁴³Adam Roberts et al., *Hierarchical Latent Vector Models for Music*, 2018. [42]

- Le *GAN* applicate alla musica, pur meno diffuse rispetto agli altri domini, trovano applicazione nella generazione di pattern ritmici e texture sonore. Un esempio significativo è *MidiNet*⁴⁴, che combina reti convoluzionali con *GAN* per generare melodie coerenti condizionate su una sequenza armonica.

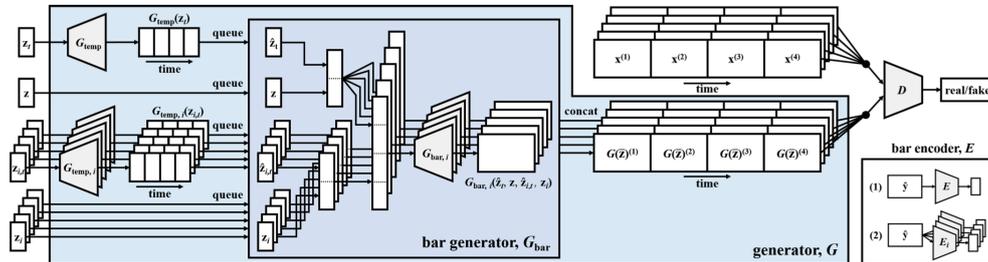


Figura 2.25: Funzionamento dei GAN nella generazione musicale.

- I modelli di diffusione, come *AudioLDM*⁴⁵ sfruttano rappresentazioni audio latenti per generare musica partendo da prompt testuali. L'audio viene rappresentato come spettrogramma latente, sul quale si applica il processo di denoising in maniera simile alla generazione di immagini. Il decoder finale ricostruisce la waveform partendo da questi spettrogrammi. Questo approccio consente una maggiore coerenza semantica tra prompt e musica risultante, permettendo quindi la generazione musicale basata su un testo.

2.5.2 Funzionamento dei Modelli per la Generazione di Effetti Sonori

La generazione automatica di effetti sonori rappresenta un sottoinsieme affine alla generazione di musica, ma con obiettivi legati alla produzione audiovisiva, ambientale o immersiva.

Un approccio comune prevede la generazione di uno spettrogramma, che viene successivamente trasformato in segnale audio tramite algoritmi di ricostruzione (come Griffin-Lim) o vocoder neurali. Il modello *DDSP*⁴⁶ introduce un paradigma innovativo: anziché operare su rappresentazioni spettrografiche passive, la rete neurale genera direttamente parametri di sintesi come armoniche, frequenza fondamentale e riverbero, consentendo una sintesi musicale controllata.

Invece, per quanto riguarda la generazione di suoni diegetici (passi, rumori ambientali, effetti) a partire da input video o testuali, sistemi come *SoundSpaces* e *Soundify* utilizzano reti neurali multimodali per collegare informazioni visive a eventi sonori coerenti. Questa tecnica trova applicazione nei videogiochi, nel cinema e nella realtà virtuale.

⁴⁴Dong et al., *MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment*, 2018. [43]

⁴⁵Liu et al., *AudioLDM: Text-to-Audio Generation with Latent Diffusion Models*, 2023. [44]

⁴⁶Jesse Engel et al., *DDSP: Differentiable Digital Signal Processing*, 2020. [45]

2.5.3 Funzionamento dei Modelli per la Generazione Vocale

La sintesi vocale (Text-to-Speech, TTS) ha compiuto notevoli progressi tecnologici nell'ultimo decennio. I tradizionali sistemi concatenativi, basati sull'accostamento di segmenti audio preregistrati, sono stati gradualmente rimpiazzati da architetture neurali, in grado di generare parlato con una naturalezza e una flessibilità significativamente superiori.

L'architettura *Tacotron 2*⁴⁷ rappresenta un esempio paradigmatico di questa evoluzione. Il modello impiega una struttura sequence-to-sequence integrata con meccanismi di attention per la generazione di spettrogrammi Mel a partire da input testuali. La conversione degli spettrogrammi in segnale audio viene successivamente realizzata mediante vocoder neurali, tra cui *WaveNet*⁴⁸.

WaveNet ha introdotto un approccio innovativo nella generazione audio, operando direttamente sulla forma d'onda attraverso una rete convoluzionale dilatata autoregressiva che processa il segnale campione per campione. Questa metodologia ha permesso di ottenere un controllo granulare sulla qualità sonora e sulle caratteristiche prosodiche del parlato sintetizzato.

Voice Cloning

L'ambito del voice cloning ha beneficiato di recenti sviluppi tecnologici che hanno ridotto drasticamente i requisiti di dati necessari per la clonazione vocale. Il sistema *VALL-E* sviluppato da Microsoft⁴⁹ esemplifica questa tendenza, dimostrando la capacità di replicare caratteristiche vocali individuali utilizzando solamente pochi secondi di materiale audio di riferimento.

Il processo si basa sull'impiego di encoder vocali che estraggono rappresentazioni latenti delle proprietà timbriche del parlante. Tali rappresentazioni vengono quindi utilizzate per guidare la sintesi di nuovo contenuto vocale, preservando le caratteristiche distintive della voce originale anche in contesti linguistici non presenti nel dataset di addestramento. Questa capacità di generalizzazione zero-shot rappresenta un avanzamento significativo rispetto ai precedenti approcci che richiedevano notevoli quantità di dati specifici per ciascun parlante.

Le implicazioni applicative di queste tecnologie spaziano dal doppiaggio automatico multilingue alla personalizzazione di assistenti vocali, fino alla ricostruzione di voci storiche per finalità documentaristiche e commemorative.

2.5.4 Comparazione IA per la Generazione Musicale

La generazione musicale tramite intelligenza artificiale rappresenta una tecnologia ancora in fase di sviluppo, infatti solamente in questo periodo sono iniziate ad essere pubblicate le prime ricerche sui confronti tra i diversi modelli disponibili. Tra queste, il lavoro di Grötschla

⁴⁷Jonathan Shen et al., *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*, 2017. [46]

⁴⁸Aaron van den Oord e Sander Dieleman, *WaveNet: A Generative Model for Raw Audio*, DeepMind 2016. [47]

⁴⁹Chengyi Wang et al., *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers*, 2023. [48]

et al. (2025)⁵⁰ propone uno dei primi benchmark sistematici basati su preferenze umane per valutare dodici modelli di generazione musicale, sia open-source sia commerciali. Usando questi dodici modelli sono state generate 6000 canzoni, che sono poi state confrontate e valutate da un pubblico di 2500 persone, i quali hanno espresso le proprie preferenze in oltre 15.000 confronti audio a coppie che condividevano lo stesso prompt in input (2.26).

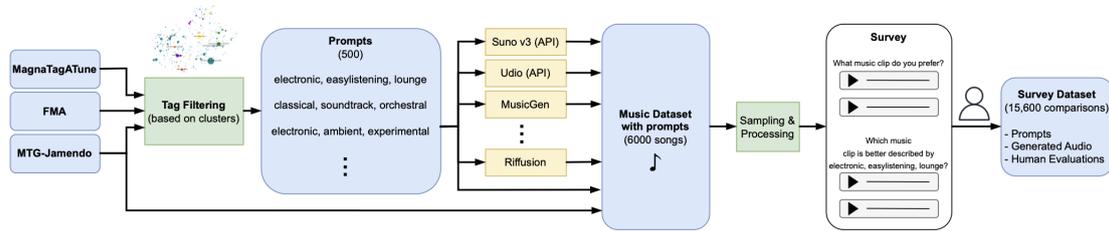


Figura 2.26: Workflow per il benchmark sui modelli di generazione musicale.

Essendo comunque la preferenza umana estremamente soggettiva, si è cercato di tradurre questi giudizi in metriche oggettive, in particolare:

- Qualità dell’allineamento tra il prompt testuale e l’audio generato.
- Qualità complessiva della musica prodotta.

In questo contesto, i 15.000 confronti sono serviti per valutare l’allineamento testo-audio e la preferenza musicale dal punto di vista umano, e per confrontare tali valutazioni con le metriche oggettive precedentemente citate.

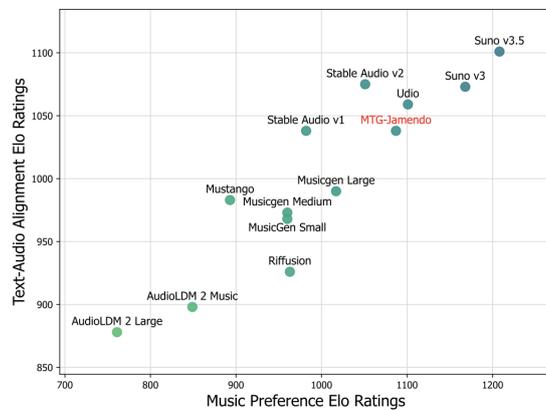


Figura 2.27: Grafico sui benchmark dei modelli di generazione musicale.

Come si può evincere dalla figura 2.27, tra i modelli confrontati, quelli che restituiscono i punteggi maggiori nei due parametri riscontrati, sono:

- *Suno v3.5*⁵¹.
- *Suno v3*.

⁵⁰Florian Grötschla et al., *Benchmarking Music Generation Models and Metrics via Human Preference Studies*, 2025. [49]

⁵¹Suno Inc., *Suno*, 2023. [50]

- *Udio*⁵².

Risulta evidente come il modello commerciale *Suno* sia attualmente quello più in linea con i giudizi umani, distinguendosi per aderenza al prompt testuale e piacevolezza percepita. Ciononostante, il campo della generazione musicale è in rapida evoluzione, e lo stesso benchmark qui analizzato è ancora in fase di sviluppo, segno di una disciplina in piena crescita e definizione.

2.5.5 Comparazione IA per la Generazione di Effetti Sonori

Il campo della generazione di effetti sonori tramite intelligenza artificiale è relativamente nuovo rispetto alla sintesi vocale e musicale, ma ha registrato sviluppi significativi negli ultimi anni. A differenza della generazione musicale e vocale, per la quale esistono benchmark sistematici consolidati, la valutazione dei modelli di generazione di effetti sonori si basa principalmente su comparazioni qualitative e su metriche di fedeltà audio. I principali modelli attualmente disponibili per la generazione di effetti sonori includono soluzioni sia commerciali che open-source. Tra i modelli più rilevanti nel panorama attuale si distinguono:

- *AudioGen (Meta)*⁵³: addestrato su effetti sonori pubblici, genera audio da prompt testuali. Questo modello fa parte della suite AudioCraft di Meta e rappresenta una delle soluzioni open-source più avanzate nel settore. AudioCraft è una base di codice unica per tutte le esigenze audio generative: musica, effetti sonori e compressione dopo l'addestramento su segnali audio grezzi.
- *ElevenLabs Sound Effects*⁵⁴: piattaforma commerciale che offre generazione di effetti sonori tramite API, con focus su qualità audio e facilità d'uso.

A differenza dei benchmark strutturati disponibili per la generazione musicale e vocale, la valutazione dei modelli di generazione di effetti sonori si basa principalmente su:

- Fedeltà al prompt: capacità del modello di generare suoni che corrispondono accuratamente alla descrizione testuale fornita.
- Qualità audio: chiarezza, assenza di artefatti e realismo del suono generato.
- Tipologia di accesso: se i servizi sono disponibili online o se è necessario utilizzare il modello in locale.
- Costo: se il modello è gratuito o meno.
- Numero di generazioni per prompt: quanti output sono forniti ogni volta che viene elaborato un input.
- Durata: lunghezza massima degli effetti sonori generabili.
- Hardware richiesto: i requisiti di sistema nel caso in cui il modello si possa utilizzare solamente in locale.
- Velocità di generazione: tempo necessario per produrre un effetto sonoro.

⁵²Udio Inc., *Udio*, 2024. [51]

⁵³Felix Kreuk, *AudioGen: Textually Guided Audio Generation*, 2023. [52]

⁵⁴ElevenLabs Inc., *ElevenLabs*, 2023. [53]

Per fornire un confronto sistematico delle caratteristiche principali, nella Tabella 2.7 sono analizzati i due modelli secondo i criteri di valutazione precedentemente identificati:

Aspetto	AudioGen	ElevenLabs
Fedeltà al prompt	Elevata	Elevata
Qualità audio	Ottima	Ottima
Tipologia di accesso	Locale	Online
Costo	Open source	A pagamento, con alcune generazioni gratuite giornaliere
Numero di generazioni per prompt	Personalizzabile	5
Durata	10 secondi	22 secondi
Hardware richiesto	GPU 16GB+	Non richiesto
Velocità	Dipende dall'hardware	Immediata (cloud)

Tabella 2.7: Confronto pratico tra *AudioGen* e *ElevenLabs*.

Come evidenziato dalla tabella, i due modelli presentano approcci complementari: AudioGen offre una soluzione gratuita per ricercatori e sviluppatori con risorse computazionali adeguate, mentre ElevenLabs fornisce un servizio commerciale immediato senza requisiti hardware specifici.

Il settore della generazione di effetti sonori presenta ancora alcune limitazioni significative:

- Dataset limitati: Meta sta rilasciando modelli AudioGen pre-addestrati che consentono la generazione di suoni ambientali e vari effetti sonori, ma la disponibilità di dataset di alta qualità per l'addestramento rimane limitata.
- Controllo granulare: la possibilità di controllare con precisione parametri specifici degli effetti sonori (intensità, durata, tonalità) è ancora in fase di sviluppo.

2.5.6 Comparazione IA per la Generazione Vocale

Per effettuare il confronto sulle intelligenze artificiali per la generazione vocale, nella funzione Text-to-Speech, verrà utilizzato un benchmark di ArtificialAnalysis⁵⁵, fonte già utilizzata nel paragrafo 2.3.2 per il confronto delle IA di generazione immagini. In particolare, sono confrontati trentanove modelli secondo le seguenti metriche:

- Qualità.
- Caratteri al secondo generati.
- Prezzo.

Verrà ora analizzato ciascuno di questi tre criteri, riportando quelli che sono i modelli che hanno ottenuto i migliori risultati.

Qualità

Come nel caso della generazione di immagini, anche per la sintesi vocale viene utilizzato un punteggio chiamato ELO per la valutazione della qualità delle generazioni. Questo punteggio ELO è assegnato dagli utenti sulla base delle preferenze da essi espresse nella sezione di “Text-to-Speech Arena”.

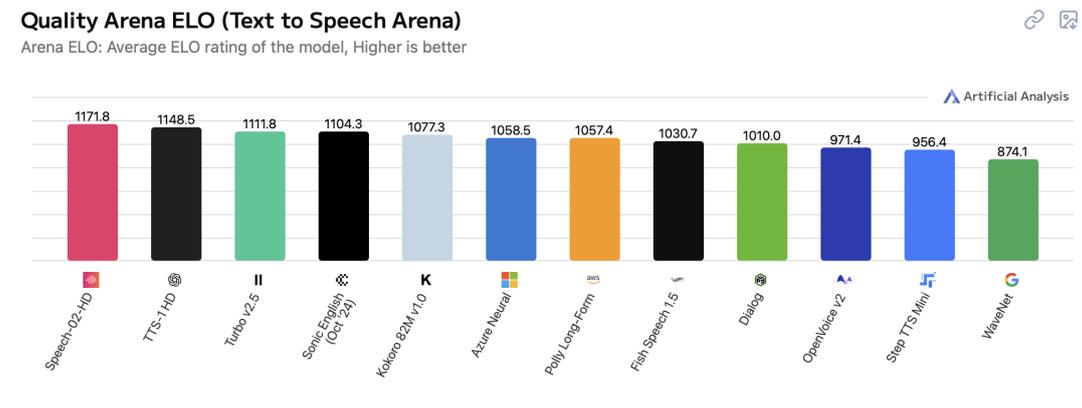


Figura 2.28: Confronto sul punteggio ELO.

⁵⁵ArtificialAnalysis Inc., *Text to Speech Models and Providers Leaderboard*, 2025. [54]

È fornita anche la percentuale di vittoria dei modelli nella “Text-to-Speech Arena”:

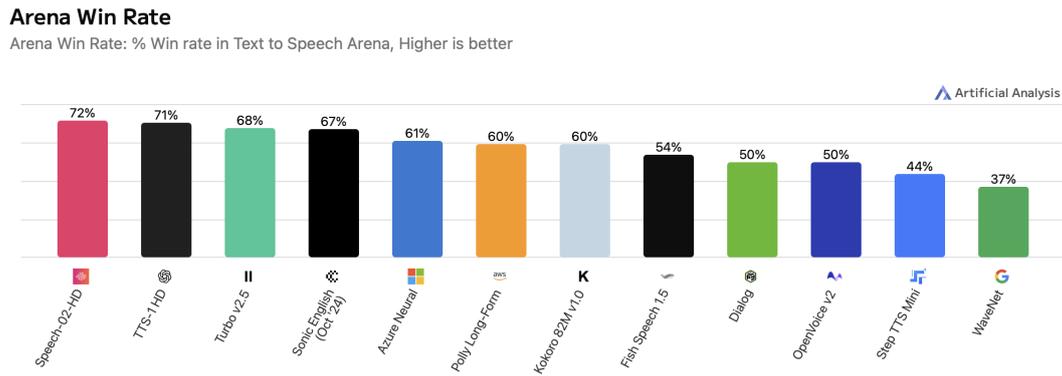


Figura 2.29: Confronto sulla percentuale di vittoria nella “Text-to-Speech Arena”.

Dai dati sulla qualità emerge come i tre modelli maggiormente performanti nelle generazioni siano:

- *Speech-02-HD* di Replicate⁵⁶.
- *TTS-1 HD* di OpenAI⁵⁷.
- *Turbo v2.5* di ElevenLabs⁵⁸.

Caratteri al secondo generati

Questo parametro di valutazione misura i caratteri processati al secondo, ed è una buona metrica per giudicare l’efficienza dei modelli. Chiaramente, maggiore è il numero di caratteri generati e maggiore sarà l’efficienza.

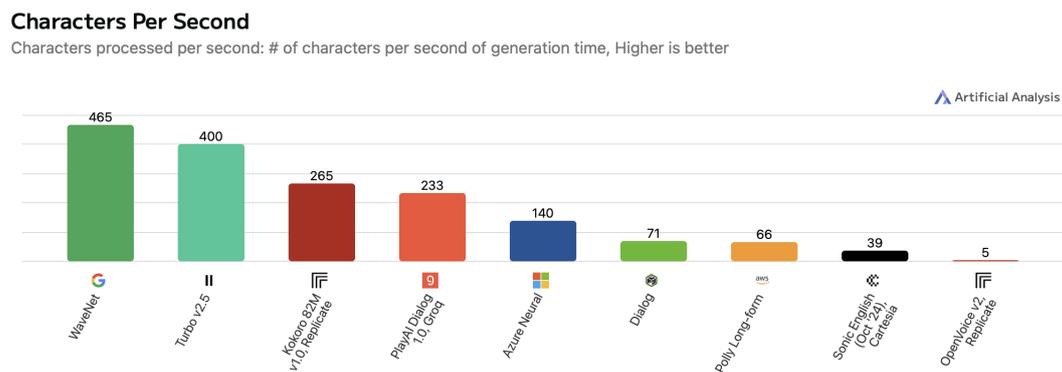


Figura 2.30: Confronto sul numero di caratteri al secondo generati.

Aggiungendo alla valutazione anche la varianza, ovvero di quanto i tempi potrebbero discostarsi dalla media, si può ottenere una valutazione sulla stabilità dell’efficienza di questi modelli.

⁵⁶Replicate, *Speech-02-HD*, 2025. [55]

⁵⁷OpenAI, *TTS-1 HD*, 2024. [56]

⁵⁸ElevenLabs Inc., *Turbo v2.5*, 2024. [57]

Characters Per Second, Variance

Characters processed per second: # of characters per second of generation time, Results by percentile, Higher is better
 — Median, Other points represent 5th, 25th, 75th, 95th Percentiles respectively

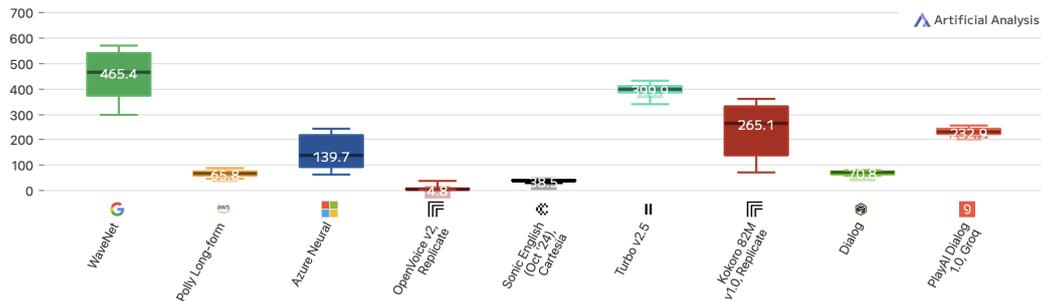


Figura 2.31: Confronto sul numero di caratteri al secondo generati comprensivo di varianza.

Da questi dati risulta che *WaveNet* di Google sia il più veloce, generando 465 caratteri al secondo, seguito da *Turbo v2.5* di ElevenLabs e da *Kokoro 82M* di Replicate. Tenendo anche conto della varianza, si evince come tra questi, il modello più stabile risulti essere *Turbo v2.5*.

Prezzo

Il prezzo misura il costo in dollari per la sintesi di un milione di caratteri, ed è un fattore cruciale in casi in cui il budget sia limitato.

Price

Price: USD per 1M characters of text, Lower is better

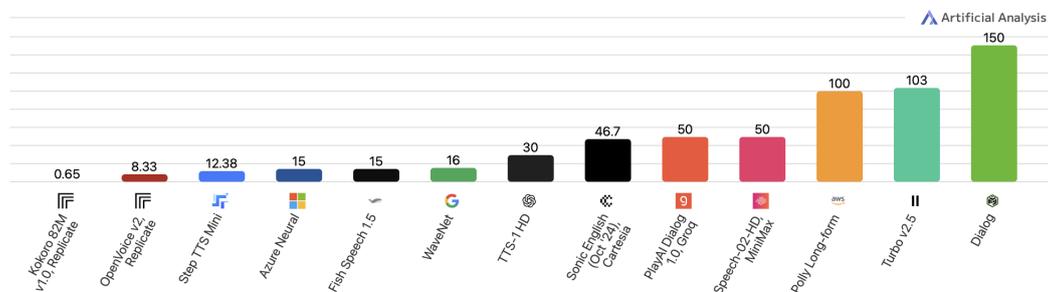


Figura 2.32: Confronto sul prezzo in dollari per la sintesi vocale di un milione di caratteri.

I modelli che risultano costare meno sono:

- *Kokoro 82M* di Replicate, con 0.65\$.
- *OpenVoice v2* di Replicate, con 8.33\$.
- *Step TTS Mini*, con 12.38\$.

Capitolo 3

Elaborato di tesi

Il presente capitolo descrive nel dettaglio il percorso progettuale e produttivo che ha condotto alla realizzazione del video finale generato con intelligenza artificiale, nato dalla collaborazione con il progetto culturale Narrazioni Parallele. L'obiettivo principale è stato quello di restituire, attraverso una narrazione audiovisiva, le emozioni, i pensieri e le immagini mentali scaturite nei partecipanti durante due distinti eventi musicali.

Per affrontare questa sfida, è stato necessario ideare un workflow alternativo a quello cinematografico tradizionale, capace di integrare le più recenti tecnologie di IA generativa in ogni fase del processo: dalla raccolta dei dati emotivi all'elaborazione dello script, dalla definizione dello stile visivo alla generazione automatizzata delle immagini e dei video, fino al montaggio finale.

Il capitolo si apre con una riflessione critica sulla scelta del metodo produttivo, analizzando vantaggi e limiti degli approcci online e locali. Segue l'analisi dei dati raccolti durante i due eventi, utilizzati per informare la narrazione e definire il contenuto visivo. Viene quindi descritta l'elaborazione dello script mediante diversi LLM (Large Language Models), la selezione dello stile visivo e la creazione dello storyboard, fino ad arrivare alla produzione del video attraverso modelli di video generation, alla generazione della musica, dei suoni e della voce, ed infine, alla fase di montaggio e valutazione del risultato.

3.1 Tipologia di Workflow

La definizione di un workflow strutturato rappresenta un prerequisito fondamentale per qualsiasi processo di generazione e elaborazione di contenuti multimediali. Nel contesto della produzione cinematografica tradizionale, tale metodologia si articola attraverso fasi ben definite che garantiscono la coerenza e la qualità del prodotto finale.

Il workflow convenzionale per la produzione cinematografica comprende le seguenti fasi sequenziali:

- Stesura dello script: definizione della struttura narrativa, dei dialoghi e delle indicazioni tecniche;
- Composizione dello storyboard: visualizzazione grafica delle sequenze attraverso schizzi e annotazioni tecniche;
- Produzione: realizzazione delle riprese secondo le specifiche definite nelle fasi precedenti;
- Post-produzione: elaborazione e montaggio del materiale acquisito per ottenere il prodotto finale;
- Export: finalizzazione del contenuto.

L'integrazione dell'intelligenza artificiale generativa in questo processo e nell'ottica di produrre un video che rievochi determinate sensazioni, richiede necessariamente un adattamento metodologico significativo. Le modifiche principali al workflow tradizionale includono:

- Fase di ricerca preliminare: prima della composizione del prompt per la generazione dello script, risulta indispensabile condurre una raccolta dati sistematica relativa alla risposta emotiva del pubblico. Nel caso specifico dei concerti di Narrazioni Parallele, questa fase comprende l'analisi delle reazioni dell'audience per derivare linee guida tematiche coerenti con le emozioni e le sensazioni del pubblico di riferimento;
- Sostituzione della fase di produzione: la tradizionale fase di produzione viene sostituita da un processo di generazione automatizzata di materiali multimediali, includendo contenuti video, audio, voiceover e colonne sonore, attraverso algoritmi di intelligenza artificiale.

Per l'implementazione delle fasi di composizione dello storyboard e generazione dei contenuti, si sono identificati due approcci metodologici principali:

- Approccio online: utilizzo di servizi online per la generazione di contenuti;
- Approccio locale: impiego di modelli open-source eseguiti su computer in locale.

La scelta tra questi approcci riveste un'importanza fondamentale per la qualità del risultato finale. Nei paragrafi successivi verranno analizzati dettagliatamente entrambi gli approcci, valutandone i vantaggi e le criticità operative.

3.1.1 Approccio Online

L'approccio online si basa sull'utilizzo di servizi di intelligenza artificiale accessibili tramite interfacce web, che possono essere sia gratuiti che a pagamento. Prima di procedere con l'a-

nalisi dettagliata del workflow basato su questo approccio, è necessario identificare e valutare le problematiche tecniche e operative che potrebbero emergere durante l’implementazione.

Durante la conferenza “AI Storytelling and Cinematography: come i modelli open source stanno rivoluzionando la pipeline cinematografica”¹, tenutasi il 27 maggio 2025 presso il Politecnico di Torino sotto la direzione di Davide Tromba, CEO e Producer degli Animoka Studios, sono state evidenziate le seguenti criticità tecniche:

- Incoerenza stilistica: inconsistenza visiva tra fotogrammi consecutivi che compromette la continuità narrativa;
- Perdita di dettagli: degradazione qualitativa nei soggetti principali o negli elementi di sfondo durante il processo generativo;
- Discontinuità narrativa: interruzioni nella coerenza logica delle sequenze complesse;
- Limitazioni temporali: durata fissa dei video generati, tipicamente limitata a cinque secondi per singola generazione;
- Costi operativi elevati: onerosità economica dei servizi online di qualità professionale.

Queste problematiche delineano sostanzialmente l’impossibilità per l’utente di mantenere un controllo preciso sui parametri di generazione, limitando significativamente la prevedibilità e la riproducibilità dei risultati.

Nel corso di questo elaborato di tesi si è ricercato un sistema metodologico per massimizzare il controllo utente sui processi generativi.

Si è prima provato ad utilizzare la tecnica del *Text-to-Video*, che permette di generare contenuti video tramite la sola immissione di un prompt testuale. In questo caso, la mancanza di controllo risulta ancora più accentuata, richiedendo prompt estremamente precisi e dettagliati. Un’altra problematica derivante da questa metodologia è l’incoerenza delle immagini mostrate: in due sequenze consecutive, è molto complicato riuscire a mantenere un soggetto completamente coerente, dato che le generazioni sono a sé stanti. Una possibile soluzione a questa problematica potrebbe essere l’ideazione di un prompt estremamente preciso che vada a descrivere dettagliatamente il soggetto, e che poi venga aggiunto ai prompt descrittivi delle varie sequenze in maniera sistematica. Questo metodo migliora la coerenza del soggetto nella maggior parte dei casi, ma non la risolve completamente: spesso bisogna comunque affidarsi a generazioni iterate prima di ottenere un risultato soddisfacente, andando quindi ad aumentare notevolmente i tempi di generazione e, soprattutto, i costi.

Ricercando metodologie migliori, lo storyboard si è rivelato uno strumento d’importanza strategica fondamentale, evolvendosi da semplice strumento di pianificazione visiva a componente integrante del processo produttivo stesso.

La soluzione ottimale identificata, infatti, consiste nell’implementazione della tecnica *Image-To-Video*, che permette all’intelligenza artificiale di generare sequenze video a partire da una o più immagini fornite dall’utente come input di riferimento. Questa metodologia, quando

¹Davide Tromba, 27 maggio 2025, *AI Storytelling and Cinematography: come i modelli open source stanno rivoluzionando la pipeline cinematografica.*, Conferenza presentata al Politecnico di Torino, Torino, Italia. [11]

combinata con servizi online che supportano la tecnica *Image-to-Image Video*, consente all'utente di definire con precisione i fotogrammi di inizio e fine di ciascuna sequenza video.

La tecnica *Image-to-Image Video* permette la selezione di due immagini chiave: la prima rappresenta il fotogramma iniziale della sequenza, mentre la seconda definisce il fotogramma finale. Questo approccio consente di creare collegamenti coerenti tra sequenze video consecutive, mitigando efficacemente le problematiche relative alla limitazione temporale di cinque secondi e all'incoerenza stilistica tra fotogrammi. Inoltre, rimane disponibile la possibilità di fornire prompt testuali per guidare l'algoritmo nella generazione della transizione tra le immagini chiave.

Utilizzando questo approccio metodologico, si possono distinguere due categorie principali di video generati:

- Video *Image-to-Image Video*: ideali per implementare transizioni controllate e passaggi precisi all'interno di sequenze narrative specifiche;
- Video *Image-To-Video*: utilizzati quando non è necessario definire un fotogramma finale specifico, permettendo maggiore libertà creativa nella generazione.

Nel secondo caso, la metodologia prevede la successiva estrazione dell'ultimo fotogramma del video generato per utilizzarlo come fotogramma iniziale della sequenza successiva, creando così un senso di continuità visiva e narrativa tra contenuti consecutivi.

Questa tipologia di approccio presenta diversi vantaggi operativi:

- Facilità d'uso: una volta completato lo storyboard, l'utente può generare contenuti senza dover gestire processi tecnici complessi, attraverso un procedimento iterativo che termina al raggiungimento del risultato desiderato;
- Velocità di generazione: i servizi online permettono tempi di generazione rapidi, eliminando la necessità di addestrare modelli computazionalmente onerosi;
- Accessibilità: non richiede hardware specializzato o competenze tecniche avanzate da parte dell'utente finale.

Tuttavia, permane una criticità significativa nella gestione dei movimenti di camera, elemento fondamentale nella produzione cinematografica professionale. Sebbene alcuni servizi online offrano opzioni per specificare il tipo di movimento di camera desiderato, l'implementazione risulta spesso inaccurata o tecnicamente inadeguata per standard professionali.

3.1.2 Approccio Locale

Nell'ambito della produzione cinematografica professionale, ogni elemento del processo produttivo è soggetto a un controllo rigoroso e a una pianificazione dettagliata. Conseguentemente, la creazione di contenuti per questo settore non può prescindere da un controllo granulare sui parametri di generazione dell'intelligenza artificiale.

Per questa ragione, l'utilizzo esclusivo di servizi online risulta spesso inadeguato per la produzione di contenuti cinematografici di qualità professionale. Sebbene una soluzione definitiva a questa problematica non sia ancora disponibile nel panorama tecnologico attuale, esistono metodologie che si avvicinano ad un controllo ottimale. Come evidenziato da Davide

Tromba², i recenti sviluppi tecnologici indicano una preferenza crescente verso approcci locali per la generazione di contenuti, sfruttando modelli open-source e operando direttamente sui dataset di addestramento.

Le strategie principali per ottenere un maggiore controllo sui contenuti generati includono:

- Creazione di dataset dedicati: sviluppo di dataset personalizzati basati sugli asset specifici del progetto e sullo stile visivo desiderato, garantendo maggiore coerenza e fedeltà alla visione creativa originale;
- Fine-tuning del modello principale: implementazione di processi di addestramento specifici per adattare la generazione ai vincoli narrativi ed estetici del progetto, migliorando significativamente la qualità e la prevedibilità dei risultati;
- Implementazione di LoRA (Low-Rank Adaptation): sviluppo di moduli computazionalmente leggeri e specifici per personalizzare rapidamente il comportamento del modello IA, ottimizzando l'utilizzo delle risorse computazionali e accelerando il processo produttivo.

La tecnica LoRA rappresenta un metodo avanzato di personalizzazione dei modelli di intelligenza artificiale, ottenuto attraverso processi di addestramento mirati su dataset personalizzati e adattamenti algoritmici specifici. Questo sistema si caratterizza per due proprietà fondamentali:

- Training mirato: addestramento su dataset personalizzati sviluppati specificatamente per il progetto, contenenti esempi rappresentativi dello stile e del contenuto desiderato;
- Adattamento specifico: implementazione di modifiche mirate al modello base per ottimizzare le prestazioni su compiti specifici, mantenendo una dimensione computazionale ridotta rispetto al modello originale.

Queste caratteristiche rendono la tecnologia LoRA probabilmente il punto di forza più rilevante nelle pipeline open-source per la produzione basata su intelligenza artificiale, grazie alla sua flessibilità operativa e capacità di adattamento a contesti specifici.

L'efficacia di questo approccio è dimostrata da diversi progetti cinematografici che hanno ottenuto riconoscimenti internazionali, come il cortometraggio "To Dear Me" di Gisele Tong³, che verrà analizzato dettagliatamente come caso studio nel paragrafo 3.4.2.

Tuttavia, l'implementazione della tecnologia LoRA presenta anche criticità operative significative:

- Requisiti hardware elevati: necessità di 60-80 GB di VRAM e configurazioni multi-GPU per un addestramento efficace;
- Tempistiche di addestramento estese: il processo può richiedere da diversi giorni a settimane, soprattutto per la generazione video;

²Davide Tromba, 27 maggio 2025, *AI Storytelling and Cinematography: come i modelli open source stanno rivoluzionando la pipeline cinematografica.*, Conferenza presentata al Politecnico di Torino, Torino, Italia. [11]

³Gisele Tong, *To Dear Me*, 2024. [58]

- Competenze specialistiche: richiede conoscenze tecniche avanzate in machine learning e gestione di infrastrutture computazionali.

Un'ulteriore criticità è rappresentata dalla necessità di sviluppare dataset personalizzati, processo che richiede competenze specifiche nella raccolta dei dati e tempi di sviluppo considerevoli, risultando tuttavia fondamentale per il successo dell'approccio locale.

3.1.3 Scelta del Workflow

A seguito di un'analisi comparativa approfondita e di valutazioni di fattibilità sui due approcci metodologici precedentemente esposti, la realizzazione di questo elaborato di tesi seguirà un approccio online, in particolare tramite le tecniche di *Image-to-Video* e *Image-to-Image Video*, pur essendo consapevoli che tale scelta comporterà necessariamente un compromesso in termini di qualità finale del prodotto.

Le risorse computazionali, le competenze specialistiche e l'infrastruttura tecnologica richieste per sviluppare efficacemente un approccio locale basato su tecnologia LoRA eccedono le disponibilità del presente progetto di ricerca. L'implementazione di tale approccio richiederebbe infatti:

- Un team multidisciplinare con competenze specifiche in machine learning, computer vision e produzione cinematografica;
- Infrastrutture hardware specializzate con costi operativi elevati;
- Tempi di sviluppo e addestramento estesi.

Pertanto, l'approccio online rappresenta la soluzione più pragmatica per dimostrare la fattibilità metodologica del workflow proposto, pur riconoscendo le limitazioni qualitative intrinseche che caratterizzano questa scelta operativa.

3.2 Analisi dei Dati

Prima di procedere con l'analisi dei dati, si è resa necessaria la ricerca e l'adozione della soluzione più adeguata tra le tecnologie disponibili sul mercato per l'acquisizione dei dati, con particolare attenzione agli strumenti per la somministrazione di sondaggi live e al coinvolgimento attivo degli spettatori. L'applicazione scelta avrebbe consentito la formulazione di domande in parte orientate agli interessi specifici di Narrazioni Parallele e in parte mirate a raccogliere dati utili per il conseguimento degli obiettivi dell'elaborato di tesi.

A seguito di un'analisi comparativa delle diverse piattaforme disponibili, che ha evidenziato una generale uniformità nei servizi offerti, si è optato per l'utilizzo di Mentimeter⁴, una delle soluzioni più consolidate e diffuse nel settore. Si tratta di una piattaforma digitale che consente di rendere le presentazioni interattive e coinvolgenti, grazie all'integrazione di elementi quali sondaggi, quiz e nuvole di parole. I partecipanti, in maniera del tutto anonima, possono intervenire in tempo reale utilizzando qualsiasi dispositivo connesso ad internet, accedendo tramite un codice univoco che elimina la necessità di registrazioni complesse.

La caratteristica che ha reso questa applicazione particolarmente adatta al progetto è la possibilità di visualizzare in tempo reale i risultati delle interazioni sotto forma di rappresentazioni grafiche, restituendo quindi un feedback chiaro e immediato al pubblico.

3.2.1 Resoconto Survey “Una Passeggiata Creativa”

Il primo incontro svoltosi in collaborazione con Narrazioni Parallele è stato intitolato “Una Passeggiata Creativa” ed è stato descritto nel capitolo 1.3.1.

Ad accompagnare quest'incontro, si è utilizzato Mentimeter, sfruttando in particolare le sue funzioni di sondaggi, nuvole di parole e domande a risposta aperta. Verranno ora analizzate quelle che sono state le domande e le relative risposte utili allo sviluppo di questo elaborato di tesi.

1. “Scegli un colore che rappresenta le tue emozioni in questo momento”:

Il colore è infatti uno degli strumenti più potenti a nostra disposizione per veicolare specifiche emozioni, e risulta fondamentale in questo contesto, dove l'obiettivo è appunto rievocare tramite un video le sensazioni scaturite dall'ascolto musicale.

Di seguito i risultati scaturiti da questa prima domanda:

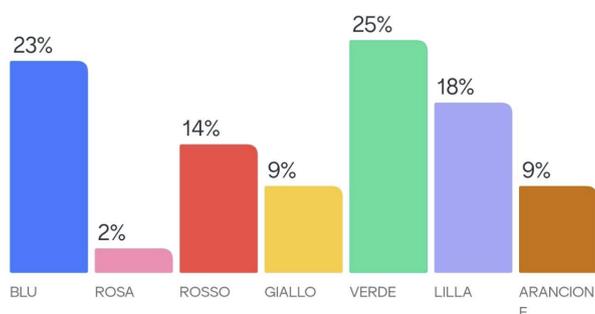


Figura 3.1: Risultati del sondaggio sui colori posto durante l'incontro.

⁴Mentimeter, *Mentimeter*, 2014. [59]

È interessante evidenziare come molte di queste risposte siano legate a specifici parti del concerto, rendendo chiaro il come ciascuna composizione evochi immagini e sensazioni diverse.

4. “Quale esperienza di vita passata ti ha ricordato l’ascolto? Cosa hai visualizzato?”: Anche questa domanda ha lo scopo di identificare possibili immagini da mostrare, accedendo a queste tramite i ricordi degli spettatori.

Anche in questo caso procederò con un elenco che racchiuda alcune delle numerose risposte:

- Una cascina con animali.
- L’infanzia.
- Il Cammino di Santiago.
- Il conservatorio.
- I giochi con gli amici durante i picnic.
- Un pomeriggio in un caffè storico di Torino.
- Una passeggiata in un viale alberato.
- L’odore delle poltrone di velluto.

In corrispondenza con la precedente domanda, si può notare come alcuni temi ricorrenti siano l’infanzia, la natura e le passeggiate.

Resoconto Google Forms successivo all’incontro

A seguito dell’incontro, è stata richiesta la disponibilità dell’audience a completare un questionario su Google Forms nei giorni successivi, attraverso il quale poter raccogliere ulteriori opinioni dei partecipanti sull’esperienza svolta.

Questo sondaggio è servito per lo più a Narrazioni Parallele, in quanto conteneva prevalentemente domande rilevanti per le loro ricerche. Ne abbiamo comunque approfittato, inserendo una domanda che ci potesse fornire un’idea su quelle che erano le aspettative del pubblico riguardo al nostro elaborato. Nello specifico, la domanda posta è stata:

- “Quale stile di video ti piacerebbe vedere nell’elaborato generato con l’Intelligenza Artificiale?”:

Questa domanda servirà più che altro a valutare ciò che il pubblico si può aspettare, e non costituirà una vera e propria base dalla quale partire per la realizzazione del video.

Tra le risposte si trovano diversi generi molto diversi tra loro, come: “Documentario”, “Realistico”, “Animato”, “Fantascienza”, “Moderno”, “Narrativo”.

3.2.2 Resoconto Survey “Song Writing Intelligenza Generativa”

Il secondo incontro svoltosi in collaborazione con Narrazioni Parallele è stato intitolato “Song Writing Intelligenza Generativa” ed è stato descritto nel capitolo 1.3.2.

Ad accompagnare quest’incontro, si è riutilizzato Mentimeter, in quanto aveva prodotto un discreto riscontro da parte del pubblico durante il primo incontro, sfruttando in particolare

le sue funzioni di sondaggi, nuvole di parole e domande a risposta aperta. Verranno ora analizzate quelle che sono state le domande e le relative risposte utili allo sviluppo di questo elaborato di tesi.

1. “Quale colore assoceresti al tema trattato oggi?”:

Seguendo la stessa struttura del survey preparato per il primo incontro, si è deciso di mantenere la domanda relativa ai colori. Tuttavia, è cambiata la modalità di visualizzazione, poiché i feedback del pubblico hanno evidenziato una preferenza per la nuvola di parole. Questo cambiamento è stato dettato principalmente da un fattore di intrattenimento, in quanto l’aggiornamento in tempo reale della nuvola di parole è risultato essere più coinvolgente rispetto ad un sondaggio.



Figura 3.3: Risultati della nuvola di parole sui colori.

Come per il primo incontro, anche in questo caso il blu risulta essere uno dei colori più scelti, seguito dall’azzurro, dal giallo e dal verde.

2. “Se la canzone fosse un viaggio, verso quale destinazione ti porterebbe?”:

In maniera concorde a quella che era la tematica dell’incontro, si è deciso di lasciare al pubblico la libertà di guidare l’esperienza stessa, e di conseguenza questa domanda è stata fatta per lasciar esprimere le persone su quello che sarebbe stato il background della canzone. Ai nostri scopi, questa domanda è servita a delineare quelle che sono le immagini mentali che la musica ha suscitato nella testa degli spettatori.



Figura 3.4: Risultati della nuvola di parole sulle “destinazioni”.

3. “Il tema del viaggio che cosa ti riporta alla memoria? Quali ricordi, desideri, film, canzoni, persone, citazioni?”:

In linea con la domanda precedente, anche questa vuole suggerire eventuali immagini da poter mostrare nel video. Essendo questa una domanda a risposta aperta, verrà delineato un elenco che riassume le principali risposte:

- “I veri viaggiatori sono coloro che partono per partire”, *Baudelaire*.
- Le pizze mangiate sulla spiaggia.
- Viaggiare in solitaria in auto, con musica e pensieri che si rincorrono e si intrecciano, un “viaggio nel viaggio” tra paesaggi reali e mentali.
- L’oceano che si infrange sugli scogli.
- Bambine che giocano nel camper, danzano, abbracciano il mondo dolcemente.
- Una bicicletta, il sole e il vento del Marocco.
- Viaggio in camper in Sardegna e la conoscenza del mio attuale gatto.

4. “Se dovessi rappresentare la canzone con una stagione, quale sarebbe?:

Questo sondaggio è stato posto con il fine di lasciar esprimere il pubblico su quella che sarebbe stata la stagione in cui avrebbero ambientato la canzone, che tradotto per i nostri scopi risulterebbe essere quella in cui si ambienterebbe il video.

Di seguito vengono riportati i risultati del sondaggio:

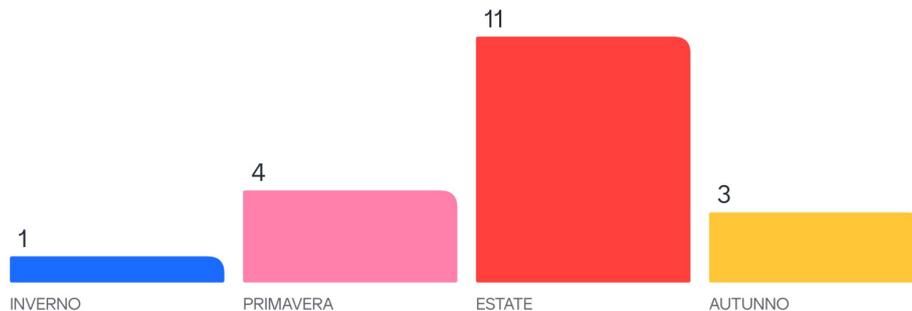


Figura 3.5: Risultati del sondaggio sulle stagioni.

In maniera coerente alle risposte precedenti, nelle quali un ipotetico viaggio al mare risultava essere la risposta più associata al tema del viaggio, la stagione più votata è stata l’estate.

5. “Il tuo contributo servirà per creare un video con l’intelligenza artificiale, quale stile di video ti piacerebbe vedere?”:

Similarmente alla domanda posta nel Google Forms al termine del primo incontro, è stato chiesto che tipologia di video il pubblico si sarebbe aspettato dall’output di quest’elaborato. Di seguito riporto un elenco delle principali risposte, per evitare ripetizioni:

- Animazione.
- Illustrazione.
- Pubblicitario.
- Docufilm.
- Astratto/Sperimentale.

Risulta evidente come siano molto diverse le aspettative sul video finale, di conseguenza, questa domanda in particolare, sarà poco rilevante nell'analisi finale dei dati che verrà svolta nel paragrafo successivo.

3.2.3 Analisi dei Dati Ottenuti

Prendendo in analisi i dati raccolti tramite i survey durante i due incontri organizzati da Narrazioni Parallele, risulta evidente come le opinioni del pubblico siano discordanti tra i due incontri. Questo è dovuto, probabilmente, al cambiamento della tipologia di musica che è stata proposta e dal ruolo del pubblico in questi eventi:

- In “Una passeggiata Creativa” si è ascoltata una rielaborazione per il quintetto di fiati della Filarmonica TRT del concerto “Quadri di un'esposizione” di Modest Mussorgsky, quindi musica classica, in cui il pubblico era ascoltatore attivo, e doveva viaggiare con la fantasia, lasciandosi trasportare dalla melodia.
- In “Song Writing: Intelligenza Naturale Generativa”, invece, si è ascoltata una composizione contemporanea di voce e chitarra co-creata con il pubblico. Quest'ultimo, essendo protagonista dell'evento, è stato sicuramente più coinvolto.

Si dovranno quindi fondere i dati dei due eventi, cercando di rimarcare i punti in comune ed evidenziando le differenze, per provare a restituire, tramite il video generato con IA, le corrette sensazioni.

Dopo aver riordinato i dati a disposizione, questi sono stati forniti come input a ChatGPT con il fine di ottenerne una corretta elaborazione.

Di seguito è riportata la risposta di ChatGPT:

Uno degli elementi più interessanti è l'uso del colore come strumento di rappresentazione emotiva. In entrambi gli incontri, il blu è stato il colore più scelto, suggerendo un senso di calma, profondità e riflessione. Tuttavia, mentre nel primo evento il verde e il lilla hanno accompagnato il blu, riflettendo un'atmosfera più onirica e sognante, nel secondo evento l'azzurro e il giallo sono emersi con maggiore forza, indicando una maggiore dinamicità e una dimensione di apertura e vitalità. Questo dato può suggerire che nel primo caso la musica abbia indotto un senso di viaggio interiore e immaginifico, mentre nel secondo caso l'esperienza è stata più concreta, proiettata verso l'esterno e associata a una dimensione esperienziale più attiva.

Le emozioni prevalenti nei due eventi si differenziano anch'esse. Nel primo incontro, le sensazioni predominanti sono state gioia, nostalgia, malinconia, pace e relax. Questo suggerisce un'esperienza emotiva stratificata, in cui la musica ha stimolato un flusso di pensieri che oscillavano tra la serenità e un senso di dolce malinconia legata al ricordo. Nel secondo incontro, invece, l'accento è stato posto più sul viaggio, inteso sia come esperienza fisica che come movimento mentale, con associazioni più concrete a luoghi, stagioni e ricordi personali. Se nel primo evento la musica era un mezzo per evocare immagini oniriche e paesaggi interiori, nel secondo evento l'esperienza musicale ha dato forma a una narrazione più strutturata, legata a esperienze vissute e a scenari reali.

Un altro punto fondamentale è l'analisi delle immagini mentali evocate. Nel primo evento, molte risposte hanno riportato visioni di paesaggi naturali e scenari fiabeschi: passeggiate nella natura, una carovana di cammelli nel deserto, mongolfiere, una casa piena di stanze musicali, una nave in mezzo al mare. Queste immagini suggeriscono una dimensione narrativa sospesa, in cui la musica fungeva da ponte per esplorare mondi immaginari. Nel secondo evento, invece, il tema del viaggio ha preso una forma più concreta, con riferimenti specifici a esperienze vissute: viaggiare in solitaria con la musica in sottofondo, il vento del Marocco, le pizze mangiate sulla spiaggia, le bambine che giocano nel camper. Qui la musica ha funzionato come catalizzatore di ricordi personali, legati non solo alla fantasia, ma anche a momenti realmente vissuti.

L'elemento della stagione conferma questa dicotomia. Nel primo incontro, non essendoci una domanda diretta sulla stagionalità, le immagini evocavano uno scenario variegato, in cui il tempo sembrava sospeso in una dimensione atemporale. Nel secondo incontro, invece, il pubblico ha scelto prevalentemente l'estate, rafforzando l'idea di un'esperienza vissuta più intensamente, legata alla luce, al calore e al movimento. Questo elemento suggerisce che la rappresentazione visiva del video finale potrebbe integrare un forte senso di stagionalità, con un'atmosfera luminosa e avvolgente, che richiami l'intensità emotiva di un viaggio in una calda giornata estiva.

Infine, l'aspettativa del pubblico nei confronti del video generato con l'intelligenza artificiale riflette una notevole varietà di visioni stilistiche. Alcuni partecipanti hanno immaginato il video come un'animazione, altri come un'illustrazione o un esperimento visivo più astratto, mentre alcuni hanno indicato un approccio documentaristico. Questa diversità riflette la molteplicità delle esperienze vissute durante gli eventi e suggerisce che il video dovrà trovare un equilibrio tra rappresentazione concreta e interpretazione simbolica delle emozioni raccolte.

Dalla fusione di questi dati emergono alcune linee guida per la costruzione del video: la presenza dominante del blu come colore guida, l'alternanza tra immagini oniriche e ricordi vissuti, il tema del viaggio come elemento di connessione tra esperienza interiore ed esteriore, e una narrazione che riesca a trasmettere la dimensione atemporale del primo evento insieme all'energia e alla concretezza del secondo. L'intelligenza artificiale dovrà quindi elaborare una sintesi visiva che non si limiti a tradurre letteralmente i dati raccolti, ma che riesca a ricreare le atmosfere e le sensazioni espresse dai partecipanti, costruendo un racconto fluido e immersivo che restituisca il senso profondo dell'esperienza musicale vissuta.

A partire da questa analisi effettuata da ChatGPT, si procederà a definire quello che sarà lo script per il video da generare.

3.3 Generazione dello Script

3.3.1 Definizione del Prompt

Prima di allegare le informazioni precedentemente ricavate dall'analisi dei dati, è necessario decidere un paio di linee guida per quella che sarà la struttura del video finale. Innanzitutto, bisogna stabilire la durata del video: come da indicazioni, il video non potrà eccedere i dieci minuti di durata, ma, considerando che l'output sarà un video con al più una voce narrante di sottofondo, imporrei come limite di durata cinque minuti.

Avendo definito questo importante fattore, il seguente sarà il prompt che verrà dato in input a tre LLM tra quelli che, nel paragrafo 2.2.2, hanno ottenuto i punteggi migliori, nello specifico DeepSeek R1, Claude 3.7 Sonnet e ChatGPT-4o (Feb 2025):

Genera uno script per un video della durata di cinque minuti, realizzato con intelligenza artificiale, che riesca a trasmettere le sensazioni vissute dal pubblico durante due distinti eventi musicali. L'obiettivo è evocare, attraverso immagini, colori e narrazione, l'esperienza emotiva di chi ha assistito ai concerti, ricreando un racconto fluido e immersivo. Non ci devono essere riferimenti espliciti ai concerti.

Il video deve alternare momenti più onirici e astratti a sequenze più concrete e realistiche, mantenendo un equilibrio tra introspezione e dinamicità. Il colore svolge un ruolo essenziale nella narrazione: il blu deve essere la tonalità dominante, suggerendo calma, profondità e riflessione. Nella prima parte, il verde e il lilla accompagnano il blu, creando un'atmosfera sospesa e sognante. Nel secondo segmento, invece, il giallo e l'azzurro emergono con maggiore intensità, rappresentando una dimensione più attiva e vitale.

Le emozioni trasmesse devono variare nel corso del video. La prima sezione si concentra su gioia, nostalgia, malinconia, pace e relax, suggerendo un viaggio interiore fatto di paesaggi immaginari e sensazioni sospese nel tempo. Nella seconda parte, il ritmo si fa più dinamico, il movimento diventa centrale e la musica si lega a esperienze vissute, trasportando lo spettatore in un viaggio più concreto, fatto di ricordi, luoghi e stagioni ben definite.

Le immagini devono essere evocative e coerenti con queste atmosfere. Nella prima parte, il video può mostrare paesaggi naturali e fiabeschi: una nave in mezzo al mare, mongolfiere che fluttuano nel cielo, carovane nel deserto o case musicali che si aprono in mondi nascosti. Nella seconda parte, la narrazione visiva si sposta su elementi più tangibili, come il vento del Marocco, viaggi in solitaria con la musica in sottofondo, spiagge estive illuminate dalla luce dorata, bambine che giocano in un camper. La stagionalità ha un ruolo centrale nella costruzione del racconto: se la prima parte è caratterizzata da un tempo sospeso e indefinito, la seconda richiama fortemente l'estate, con il calore, la luce intensa e il senso di libertà che essa porta con sé.

Il ritmo deve alternare sequenze più lente e contemplative a momenti più dinamici e coinvolgenti, costruendo un crescendo emozionale che accompagni lo spettatore lungo tutto il viaggio. La narrazione può essere supportata da una voce fuori campo o da testi su schermo, ma senza mai sovrastare la forza evocativa delle immagini.

L'obiettivo finale è ricreare nel pubblico le stesse emozioni provate durante gli eventi musicali, trasformando dati ed esperienze raccolte in un flusso visivo e sonoro che restituisca il senso profondo di quei momenti, lasciando spazio all'interpretazione personale e all'immaginazione.

Verranno, in seguito, analizzati e comparati i tre output generati, con l'obiettivo di scegliere quello che risponda alle esigenze del progetto.

3.3.2 DeepSeek R1

DeepSeek R1⁵ è un modello di intelligenza artificiale sviluppato dalla startup cinese DeepSeek, specializzato in compiti che richiedono capacità di ragionamento avanzato, come la risoluzione di problemi matematici, la programmazione e l'inferenza nel linguaggio naturale.

Il modello si basa su un'architettura LLM (Large Language Model) open source, simile a quella utilizzata da altri sistemi avanzati come GPT-4. Tuttavia, una delle principali innovazioni di DeepSeek è l'uso di tecniche di apprendimento per rinforzo e un approccio di ragionamento passo-passo, che consente al modello di affrontare compiti complessi con maggiore efficienza e accuratezza. Questo approccio riduce significativamente i costi di sviluppo e quelli operativi rispetto ai modelli della concorrenza.

Il prompt definito nel paragrafo 3.5.1 è stato immesso come input in DeepSeek R1, ottenendo questa risposta:

- Titolo: "Due note, un respiro"
- Durata: 5 minuti
- Struttura: 2 atti (2 minuti e 30 secondi ciascuno), transizioni accelerate ma fluide

Atto 1 - Sospesi nel Blu (0:00 - 2:30)

- Emozioni: Nostalgia, Malinconia, Transizione verso la speranza
- Colori: Blu notte → Smeraldo → Lilla
- Musica: Pianoforte/archi, crescendo progressivo
- Sequenze:
 1. 0:00 - 0:25
 - Visuale: Nebulosa blu → Oceano con nave antica (vele lilla)
 - Voce fuori campo: *"C'è un luogo dove il tempo respira, e la musica è un sogno che non si spegne."*
 - Transizione: Nave → mongolfiera in 10 secondi (acqua diventa cielo)
 2. 0:25 - 1:10

⁵DeepSeek AI, *DeepSeek-R1*, 2024. [60]

- Visuale: Valli smeraldo/nuvole lilla, mongolfiere in slow-motion
- Musica: Violoncello + vento (taglia campane lontane)
- Simbolo: Bambina (comparsa rapida) raccoglie stella cadente

3. 1:10 - 1:50

- Visuale: Carovana nel deserto → Tende rivelano mondi in miniatura (solo cristalli e città sottomarine)
- Transizione: Stella → chiave di violino in 5 secondi

4. 1:50 - 2:30

- Climax: Porta nel deserto si apre → cascata di note blu/verdi
- Musica: Accelerazione più marcata
- Frame finale: Occhi della bambina si chiudono (nero per 2 secondi)

Atto 2 - Risveglio nel Giallo (2:30 - 5:00)

- Emozioni: Euforia, Libertà, Connessione tra passato/presente
- Colori: Giallo oro → Azzurro → Arancione
- Musica: Percussioni/chitarre elettriche, finale etereo
- Sequenze:

1. 2:30 - 3:10

- Visuale: Occhi della bambina → sole giallo/campo di grano (tappeti berberi in cielo)
- Voce fuori campo: *“Ma il sogno sapeva di sale, di polvere che brucia, di passi senza ritorno.”*
- Transizione: Petali → sabbia in 5 secondi

2. 3:10 - 3:50

- Scene dinamiche:
 - * Moto sulla costa (ripresa drone, 10 secondi)
 - * Ragazze nel camper (5 secondi, risate brevi)
 - * Fuochi d’artificio → sincrono con batteria (10 secondi)
- Montaggio: Tagli rapidi su ritmo percussioni

3. 3:50 - 4:20

- Flashback: Bambina (adolescente) scrive in stanza con dischi
- Specchi: Fusione deserto-spiaggia (5 secondi)
- Colore: Giallo → arancione in dissolvenza

4. 4:20 - 5:00

- Finale:
 - * Bambina cammina verso faro (fusione deserto-mare, 15 secondi)
 - * Faro acceso → vortice di colori (10 secondi)
 - * Voce finale: *“E così, la musica non finì mai. Divenne il vento, il respiro, il posto in cui ritrovarsi.”*
- Frame conclusivo: Sabbia e note musicali (5 secondi)

3.3.3 Claude 3.7 Sonnet

Claude 3.7 Sonnet⁶ è un avanzato modello di intelligenza artificiale sviluppato da Anthropic ed è stato il primo modello di “ragionamento ibrido” sul mercato, capace di fornire risposte immediate o di intraprendere un processo di riflessione più approfondito. Al suo interno integra, quindi, sia le funzionalità di un LLM tradizionale che capacità avanzate di ragionamento.

Il modello è accessibile attraverso diversi piani: Free, Pro (costo di 18 euro al mese) e Team (costo di 28 euro al mese). La modalità di “pensiero esteso” è disponibile solo per gli abbonati a pagamento. A differenziare i piani è per lo più un incremento dei limiti di utilizzo.

Il prompt definito nel paragrafo 3.5.1 è stato immesso come input in Claude 3.7 Sonnet, ottenendo questa risposta:

- Titolo: “Risonanze”
- Durata: 5 minuti
- Struttura: 2 parti (circa 2 minuti e 30 secondi ciascuna), con transizione di collegamento

Atto 1 - Mondi Interiori

- Emozioni: Gioia, Nostalgia, Malinconia, Pace, Relax
- Colori Dominanti: Blu profondo, Verde, Lilla
- Musica: Eterea, melodie sospese, strumenti acustici
- Sequenze:
 1. 00:00 - 00:15
 - Visuale: Schermo blu profondo che si dissolve lentamente. Titolo con lettere fluttuanti.
 - Voce fuori campo: “La musica non è solo ciò che ascoltiamo, ma ciò che diventiamo quando l’ascoltiamo.”

⁶Anthropic, *Claude 3.7 Sonnet*, 2024. [61]

- Transizione: Note musicali che si trasformano in farfalle di luce.
2. 00:15 - 00:45
- Visuale: Una nave solitaria naviga su un mare blu intenso. Le onde si muovono lentamente.
 - Voce fuori campo: “Ciascuno di noi, una nave solitaria, eppure tutti uniti dallo stesso orizzonte musicale.”
 - Transizione: Le onde si trasformano in nuvole blu-lilla.
3. 00:45 - 1:15
- Visuale: Mongolfiere colorate (blu, verde e lilla) si alzano nel cielo al tramonto.
 - Voce fuori campo: “Come se ogni nota ci sollevasse un po’ più in alto, verso un luogo dove la nostalgia e la gioia danzano insieme.”
 - Dettaglio: Figura solitaria che suona un violoncello dentro una mongolfiera.
4. 1:15 - 1:45
- Visuale: Un deserto blu al chiaro di luna. Una carovana di figure oniriche avanza.
 - Voce fuori campo: “La musica ci accompagna attraverso le dune del tempo, una carovana di anime in cerca di un’oasi di bellezza.”
 - Simbolismo: Le figure lasciano impronte luminose che diventano spartiti musicali.
5. 1:45 - 2:30
- Visuale: Una piccola casa blu a forma di pianoforte in mezzo a un prato verde-lilla.
 - Voce fuori campo: “Ogni canzone è una stanza da esplorare, un rifugio dove sentirsi al sicuro.”
 - Climax: La casa si apre come un libro pop-up, rivelando un mondo in miniatura.
 - Transizione: Il mondo in miniatura si dissolve in particelle di luce blu che diventano uccelli volanti.

Transizione

1. 2:30 - 2:45
- Visuale: Gli uccelli blu volano sempre più in alto.
 - Colori: Cambio graduale di palette. Il blu rimane, ma appare il giallo luminoso e l’azzurro chiaro.
 - Movimento: Un soffio di vento attraversa lo schermo, portando foglie gialle e azzurre.

Seconda Parte - Viaggi Tangibili

- Emozioni: Vitalità, Energia, Libertà, Calore, Condivisione
- Colori Dominanti: Blu con giallo luminoso e azzurro chiaro
- Musica: Ritmi più definiti, dinamici ed energici
- Sequenze:
 1. 2:45 - 3:15
 - Visuale: Un deserto sotto un cielo azzurro intenso. Bandiere colorate sventolano.
 - Voce fuori campo: “Come il vento caldo, la musica ha portato con sé profumi, colori e memorie di terre lontane.”
 - Ritmo: Movimento dinamico, volti sorridenti, mani che battono, corpi che danzano.
 2. 3:15 - 3:45
 - Visuale: Una strada costiera al tramonto, mare azzurro e campi gialli.
 - Voce fuori campo: “Ciascuno percorre un viaggio personale. Come guidare su una strada costiera, con la musica che amplifica ogni emozione.”
 - Simbolismo: Una figura guida una decappottabile blu, lasciandosi dietro note musicali.
 3. 3:45 - 4:15
 - Visuale: Una spiaggia dorata sotto un cielo azzurro intenso.
 - Voce fuori campo: “L’estate ha trovato rifugio in questa musica. Calda, vibrante, piena di vita.”
 - Movimento: Persone sulla spiaggia che ballano, ridono, si abbracciano.
 4. 4:15 - 4:50
 - Visuale: Un vecchio camper blu su una scogliera al tramonto.
 - Voce fuori campo: “Abbiamo riempito gli spazi vuoti dentro di noi con momenti di pura, incontenibile gioia.”
 - Dettagli: Bambine giocano con strumenti musicali, adulti seduti intorno a un falò.

Sequenza Finale

1. 4:50 - 5:00
 - Visuale: Scene precedenti si sovrappongono, caleidoscopio di immagini.

- Voce fuori campo: “La musica dal vivo è un ponte tra mondi. Tra il tangibile e l’etereo.”
- Finale: Un grande occhio blu che si apre e chiude, trasformandosi in nota musicale.
- Ultima voce: “Non ascoltiamo solo la musica. La viviamo. La diventiamo.”
- Titoli di coda: “FINE” in lettere luminose che si dissolvono.

3.3.4 ChatGPT-4o (Feb 2025)

GPT-4o⁷ (dove “o” sta per “omni”) è un modello multimodale e multilingue, capace di elaborare e generare testo, immagini e audio sviluppato da OpenAI, ed è un’evoluzione della serie di modelli GPT (Generative Pre-trained Transformer).

OpenAI ha reso GPT-4o accessibile gratuitamente a tutti gli utenti, con limiti di utilizzo che possono variare in base alla domanda e al livello di abbonamento. Gli abbonati a ChatGPT Plus (costo di 20 dollari al mese) beneficiano di limiti di utilizzo più elevati rispetto agli utenti gratuiti, hanno accesso ai modelli di reasoning (GPT-o3-mini) e accesso limitato alla generazione di video di SoraAI. Gli abbonati a ChatGPT Pro (costo di 200 dollari al mese) hanno accesso illimitato a tutti i modelli, alla deep research e accesso esteso alla generazione di video di SoraAI.

Il prompt definito nel paragrafo 3.5.1 è stato immesso come input in ChatGPT-4o (Feb 2025), ottenendo questa risposta:

- Titolo: “Il Suono del Ricordo”.
- Durata: 5 minuti.
- Struttura: 2 parti principali con introduzione e conclusione.

Introduzione

- 0:00 - 0:40
 - Visuale: Un cielo notturno sfumato di blu profondo, punteggiato da stelle che pulsano al ritmo di un’eco musicale lontana. Una nave fluttua su un mare di note luminose.
 - Voce fuori campo: “Ogni suono racchiude un ricordo. Ogni nota si trasforma in immagine, e noi viaggiamo dentro la musica.”
 - Transizione: Dal mare emergono onde di luce verde e lilla, che si trasformano in mongolfiere fluttuanti.

Prima Parte - Il Viaggio Interiore

⁷OpenAI, *Introduction to GPT-4o*, 2024. [62]

- 0:40 - 2:20
 - Musica: Toni eterei, arpeggi sognanti, eco lontani.
 - Visuale: Una carovana avanza nel deserto, il vento solleva polvere dorata. Le tende si aprono rivelando stanze musicali, ogni strumento suona da solo, tessendo un'armonia sospesa.
 - Voce fuori campo: “Ascoltare un concerto non è solo sentire... è perdersi in mondi nascosti, in stanze di ricordi.”
 - Visuale: Una casa in mezzo al nulla, le finestre si aprono e al loro interno si vedono paesaggi in continuo mutamento: foreste color lavanda, laghi scintillanti, aurore boreali che danzano nel cielo.
 - Musica: Un crescendo delicato, una melodia che richiama nostalgia e pace.
 - Visuale: Un volto di profilo immerso in luce blu, gli occhi chiusi. Dal suo respiro si sprigionano onde di colore, che si dissolvono nell'aria.

Seconda Parte - Il Ricordo Reale

- 2:20 - 4:20
 - Musica: Ritmo più incalzante, strumenti acustici mescolati a suoni elettronici, battiti che pulsano come un cuore in corsa.
 - Visuale: Un vento caldo attraversa le dune del Marocco, spostando i granelli di sabbia dorata. Il sole si riflette sugli occhi di un viaggiatore solitario con le cuffie.
 - Visuale: Un camper in movimento, dentro una bambina ride mentre il vento agita le tende. La strada si apre verso l'infinito, l'orizzonte è un misto di giallo e azzurro acceso.
 - Voce fuori campo: “E poi arriva il momento in cui il sogno diventa reale. La musica si mescola ai luoghi, ai gesti, agli istanti vissuti.”
 - Visuale: Spiagge assolate, onde che si infrangono, amici che corrono sulla sabbia. Il cielo si tinge di rosso mentre il giorno sfuma nella sera.
 - Musica: Un crescendo energico, note che esplodono come scintille nel buio.

Conclusione

- 4:20 - 5:00
 - Visuale: Il cielo torna blu profondo, le stelle pulsano ancora. La nave si dissolve lentamente nel mare di note.
 - Musica: Le ultime note sfumano, lasciando spazio al silenzio.

- Voce fuori campo: “Ogni concerto finisce. Ma la musica... non smette mai di suonare.”
- Finale: [Fine]

3.3.5 Valutazione e Scelta dello Script

Si riporta, nella seguente tabella, un’analisi comparativa dei tre output. È, inoltre, interessante notare come, sebbene non presente nel prompt, tutti e tre gli LLM hanno suddiviso la sceneggiatura in due atti.

Caratteristica	DeepSeek R1	Claude 3.7 Sonnet	GPT-4o (Feb 2025)
Precisione timecode	Alta (segmenti molto dettagliati)	Alta (segmenti molto dettagliati)	Media (segmenti più ampi)
Approccio cromatico	Contrasto marcato tra le due parti	Transizione graduale, blu come elemento costante	Approccio semplificato, meno dettagliato
Simbolismo	Complesso e trasformativo	Esplicito e connesso alla voce narrante	Essenziale e diretto
Elementi onirici	Nave, mongolfiere, carovana, casa musicale	Nave, mongolfiere, casa-pianoforte, mondo in miniatura	Nave su mare di note, stanze musicali, casa con paesaggi mutevoli
Elementi concreti	Moto, ragazze nel camper, diario, faro	Strada costiera, camper, spiaggia, falò	Vento del Marocco, camper, spiagge, concerti
Gestione transizione	Netta (occhi che si chiudono/riaprono)	Graduale (uccelli che volano, cambio di prospettiva)	Sottintesa (ritorno agli elementi iniziali)
Voce narrante	Poetica e simbolica	Riflessiva e collettiva	Concisa e metaforica
Dimensione emotiva	Individuale e introspettiva	Collettiva e condivisa	Universale e poetica

Tabella 3.1: Confronto tra Prompt di DeepSeek R1, Claude 3.7 Sonnet e GPT-4o.

Dopo un'attenta analisi, si può evincere come tra i tre LLM, GPT-4o (Feb 2025) sia quello che ha restituito il risultato peggiore:

- Non è stato preciso nella definizione temporale delle scene, prendendosi ampi margini di tempo in cui elencava le varie inquadrature.
- L'approccio al colore è stato meno articolato e curato.
- La connessione tra i due eventi non risulta essere esplicita.

Di conseguenza, anche visti gli ottimi output prodotti da DeepSeek R1 e Claude 3.7 Sonnet, si escluderà lo script elaborato da GPT-4o (Feb 2025).

Dover scegliere fra gli output restanti risulta essere una questione molto soggettiva, in quanto entrambi gli LLM hanno prodotto un ottimo script. Dovendo, però, effettuare una scelta, sarà scelto quello generato da Claude 3.7 Sonnet:

- Si ritiene che questo script riesca a bilanciare meglio l'aspetto emotivo con quello narrativo, rendendo più esplicito il legame fra le immagini mostrate e le sensazioni che l'audience ha provato durante i concerti.
- Tra i tre output è l'unico che parla di questi eventi come un'esperienza condivisa, riuscendo a catturare meglio la dimensione sociale di questi concerti.
- La transizione è meno netta di quella prodotta da DeepSeek R1, mantenendo un senso maggiore di continuità tra le due scene.
- Mantiene il blu come colore principale dell'intero video, ma integra in modo efficace anche le altre palette cromatiche.
- Rispetto all'output generato da DeepSeek R1, che mostrava immagini astratte e transizioni tra queste molto creative, Claude 3.7 Sonnet ha bilanciato meglio immagini reali con immagini oniriche, rispettando meglio la richiesta nel prompt di "alternare momenti più onirici e astratti a sequenze più concrete e realistiche".

Alla luce della presente analisi, lo script del paragrafo 3.5.3 sarà, quindi, quello che verrà utilizzato per la creazione del video.

3.4 Generazione dello Storyboard

Dopo aver definito la struttura narrativa del video attraverso l'analisi dei dati ed effettuato la generazione dello script, questo capitolo si concentra sulla scelta dello stile visivo più adatto a trasmettere le emozioni e le sensazioni emerse durante gli eventi analizzati. La definizione dell'estetica del video è un passaggio fondamentale, in quanto determina il modo in cui il pubblico percepirà l'opera e interagirà con essa a livello emotivo.

Si partirà con un'analisi delle diverse opzioni stilistiche, valutando come ciascuna possa influenzare l'impatto emozionale del video. Sarà poi discusso il fenomeno "Uncanny Valley", evidenziando le difficoltà che emergono nella rappresentazione realistica generata dall'IA e le ragioni per cui si è scelto di orientarsi verso un'estetica più stilizzata. Successivamente, verranno presentati alcuni casi studio rilevanti, tra cui il cortometraggio "To Dear Me" di Gisele Tong e il segmento "Corvi" dal film "Dreams" di Akira Kurosawa, per trarre ispirazione dall'uso espressivo dell'animazione e del colore.

Una volta individuato l'approccio visivo più efficace, il capitolo esaminerà le principali tecnologie di *image generation* disponibili, confrontando diversi strumenti per determinare quale sia il più adatto alla realizzazione del progetto. Infine, verrà illustrato il processo di creazione dello storyboard, che fungerà da guida per la successiva fase di generazione del video.

3.4.1 Scelta dello stile visivo

La scelta dello stile visivo del video da generare per mezzo dell'IA è una decisione cardine nel percorso creativo. Analizzando le aspettative dell'audience che ha partecipato ai due survey a seguito dei rispetti concerti svolti in eventi organizzati da "Narrazioni Parallele", risulta evidente come ci sia disaccordo su quale debba essere lo stile: alcune persone hanno immaginato uno stile documentaristico, con immagini reali, altre un docufilm, altre ancora un cortometraggio animato.

Di fronte ad una così ampia gamma di tipologie di rappresentazione, risulta necessario operare una ricerca al fine di individuare la forma artistica che più si addica al fine di trasmettere correttamente le emozioni elencate nei survey dal pubblico.

Valutando le possibili scelte artistiche, la prima che si andrà ad analizzare è quella che porta ad un'estetica fotorealistica all'interno del video.

Questa scelta porterebbe con sé una serie di vantaggi:

- Una rappresentazione fotorealistica contribuisce ad un maggiore livello di immersione dello spettatore. Infatti, l'utilizzo di ambienti e personaggi reali, può rafforzare il senso di "presence" e rendere la visione del video un'esperienza meno lontana dalla vita quotidiana, riducendo, quindi, la distanza tra il contenuto audiovisivo e il vissuto dello spettatore.
- Un altro vantaggio fornito da un'estetica realistica è la possibilità di comunicare emozioni in maniera diretta, tramite il linguaggio del corpo dei personaggi e delle espressioni facciali.

Quest'analisi, però, non può essere effettuata senza considerare il mezzo tramite il quale si andrebbe a creare il video effettivo. Infatti, quanto appena descritto nei vantaggi della

scelta di una rappresentazione realistica, non tiene conto che questo realismo non è, appunto, “reale”, ma generato artificialmente.

Questo fattore cambia totalmente il punto di vista sull’argomento, in quanto le immagini reali che contribuirebbero all’aumento dell’immersione dello spettatore, in questo contesto rischierebbero di diminuirla.

3.4.2 Il Fenomeno dell’“Uncanny Valley”

È stato dimostrato che lo spettatore prova disagio quando vede personaggi quasi realistici, ma con dettagli impercettibilmente innaturali. Questo studio è alla base del fenomeno chiamato “Uncanny Valley”⁸, che fu teorizzato nel 1970 da Masahiro Mori, ed è stato successivamente osservato in diversi contesti, dai robot umanoidi ai videogiochi, dove, appunto, una mancanza di totale realistica dei personaggi può causare nello spettatore/giocatore una sensazione di inquietudine che conduce ad un distacco dall’esperienza di gioco, compromettendo, di conseguenza, l’immersione.

Masahiro Mori rappresenta graficamente la relazione tra il grado di somiglianza di un oggetto a un essere umano e il livello di affinità emotiva suscitato negli osservatori.

Come evidenziato nella figura 3.6 (fino a circa il 70% di somiglianza con l’essere umano) all’aumentare della somiglianza con l’essere umano, si osserva una crescita dell’empatia provata nei confronti dell’oggetto: un robot giocattolo, ad esempio, risulta più attraente rispetto a un macchinario industriale, poiché, pur non riproducendo fedelmente le fattezze umane, ne richiama alcune caratteristiche. Tuttavia, quando la somiglianza diviene più precisa, si verifica un brusco calo della risposta emotiva positiva.

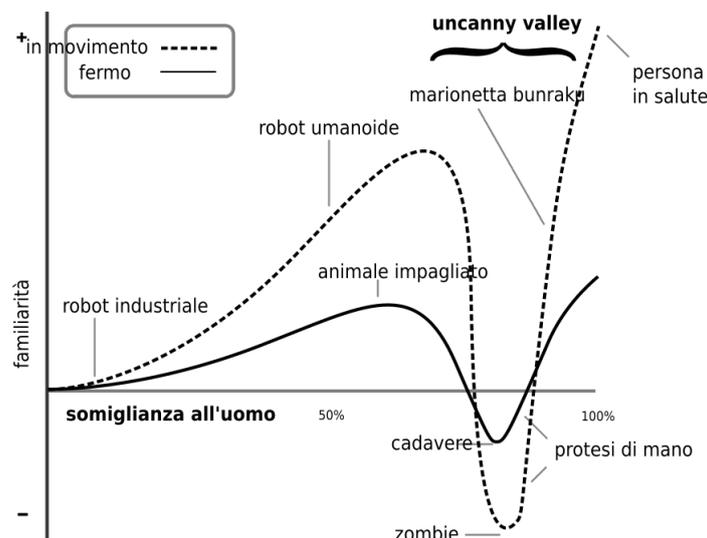


Figura 3.6: Grafico dell’Uncanny Valley.

⁸Masahiro Mori, *The Uncanny Valley*, 1970. [63]

Masahiro Mori illustra il concetto di Uncanny Valley attraverso una serie di esempi che spaziano dalla robotica industriale a oggetti inanimati, come bambole e burattini tradizionali del teatro giapponese, fino a figure immaginarie come gli zombie, che estremizzano questa percezione di inquietudine.

Nel corso degli ultimi decenni l'Uncanny Valley ha attirato l'interesse degli esperti di tecnologia, rivelandosi più attuale che mai davanti agli ultimi progressi nel campo della robotica, dell'Intelligenza Artificiale e del Machine Learning.

L'abbassamento della curva corrisponde alla sensazione di disagio ed inquietudine che gli osservatori sperimentano di fronte ad oggetti che, ad un primo sguardo, appaiono realistici, ma che rivelano la loro natura artificiale ad un'analisi più attenta. Mori osserva inoltre che tale effetto tende a intensificarsi quando l'elemento del movimento è presente.

Sebbene le tendenze di sviluppo e di incremento della qualità nelle immagini generate con IA progrediscono ogni giorno in maniera significativa, questa sensazione di disagio non è ancora del tutto superata: diversi modelli di intelligenza artificiale generativa producono immagini più che convincenti, il problema risiede nell'animazione di queste, che soprattutto con esseri umani risulta ancora acerba.

Di conseguenza, per la generazione dell'elaborato si esclude un'estetica realistica a favore di una animata.

Caso Studio: “To Dear Me” di Gisele Tong

A rafforzare questa decisione sono i festival del settore, dei quali si è parlato nel corso del secondo capitolo di questa tesi: si prende in esame l'edizione 2024 del Reply AI Film Festival⁹, dove sia il primo che il secondo classificato hanno portato prodotti animati. Il cortometraggio “One Way” di Egor Kharlamov adotta uno stile visivo ispirato al fumetto, caratterizzato da un'estetica dinamica e marcata.

Al contrario, “To Dear Me” di Gisele Tong¹⁰, vincitrice del festival, si distingue per un approccio artistico pittorico che richiama la delicatezza degli acquerelli. Questo stile visivo conferisce al film un'atmosfera onirica e intimista, perfettamente in linea con la narrazione.

In particolare, questo cortometraggio esplora un viaggio interiore attraverso la mente e i ricordi della protagonista, ponendo particolare enfasi sull'uso espressivo del colore e del suono. La palette cromatica assume un ruolo fondamentale nella costruzione dell'emotività della scena, contribuendo a evocare nostalgia, introspezione e malinconia.

Parallelamente, la colonna sonora non si limita ad accompagnare le immagini, ma diventa parte integrante della narrazione, guidando lo spettatore attraverso le emozioni del personaggio e rafforzando l'immersività dell'esperienza.

⁹Reply, *Reply AI Film Festival*, 2024. [17]

¹⁰Gisele Tong, *To Dear Me*, 2024. [58]



Figura 3.7: Frame tratto da “To Dear Me” di Gisele Tong.

Ritenendo, quindi, che uno stile animato sia il miglior modo attraverso il quale rappresentare l’elaborato, bisogna ora identificare quale tipologia di immagine e animazione utilizzare.

3.4.3 Espressionismo Come Amplificatore di Emozioni

A seguito dell’esclusione del realismo come scelta estetica per il video, a causa delle problematiche legate alla “Uncanny Valley” e ai limiti attuali delle IA nell’animazione realistica, è necessario individuare uno stile visivo che sia in grado di trasmettere efficacemente le emozioni suscitate dalla musica.

Per compiere questa scelta, sono state valutate diverse alternative stilistiche, ognuna con caratteristiche specifiche. L’analisi comparativa seguente illustra i punti di forza e le criticità di ciascun approccio.

Corrente Artistica	Caratteristica	Punti di Forza	Criticità
Surrealismo	Immagini oniriche e simboliche	Trasmette emozioni profonde, adatto a rappresentare stati d’animo astratti	Può risultare troppo distante dalla realtà per il pubblico
Cubismo	Scomposizione geometrica della realtà	Approccio innovativo, enfatizza prospettive multiple	Può rendere difficile l’identificazione emotiva

Minimalismo	Sottrazione di dettagli, forme essenziali	Elegante e pulito	Rischia di ridurre l'impatto emotivo se non ben dosato
Espressionismo	Distorsione visiva per enfatizzare le emozioni	Amplifica l'impatto emotivo, facilita l'identificazione del pubblico	Può risultare stilisticamente marcato e poco realistico

Tabella 3.2: Confronto tra correnti artistiche.

Tra queste opzioni, l'espressionismo sembra essere la corrente artistica più adeguata per questo tipo di progetto, poiché si basa sulla distorsione della realtà al fine di amplificare le sensazioni interiori: piuttosto che replicare fedelmente il mondo reale, questo stile si concentra sulla rappresentazione soggettiva dell'esperienza umana attraverso l'uso di colori intensi, prospettive alterate e forme deformate.

Queste caratteristiche aderiscono perfettamente a quelle che sono le finalità dell'elaborazione del video, che, appunto, si propone di trasmettere le emozioni che il pubblico ha provato. Inoltre, questo approccio risulta in linea con i survey proposti al pubblico, in quanto è stato esso stesso a fornire descrizioni delle proprie percezioni.

Secondo Rudolf Arnheim, uno dei più importanti studiosi della percezione artistica, l'arte non deve limitarsi a riprodurre fedelmente la realtà, ma deve trasformarla e amplificarla per rendere più evidente il suo significato emotivo. Nel suo libro "Art and Visual Perception"¹¹, Arnheim sostiene che la distorsione delle forme e dei colori consenta di comunicare concetti e stati d'animo in modo più diretto ed efficace rispetto a una rappresentazione puramente realistica.

La percezione delle immagini, infatti, non è un processo passivo, ma un'attività cognitiva complessa, in cui il cervello organizza gli stimoli visivi secondo schemi coerenti. Questo principio è alla base della Teoria della Gestalt, di cui Arnheim è stato un importante sostenitore. Secondo questa teoria, la mente umana tende a strutturare la realtà in insiemi significativi, piuttosto che percepirla come una somma di dettagli isolati.

Questa dinamica si adatta perfettamente all'estetica espressionista, in quanto le immagini stilizzate e simboliche non vengono percepite come incomplete o astratte, ma piuttosto come stimoli che il cervello riorganizza spontaneamente, attribuendo loro un significato coerente. Di conseguenza, uno spettatore che osserva una figura deformata o un paesaggio irrealista non si sofferma sulla loro incongruenza rispetto alla realtà, ma li interpreta in base alla propria esperienza emotiva, lasciandosi guidare dalle sensazioni suscitate dall'opera, piuttosto che dalla sua fedeltà alla rappresentazione del mondo reale.

¹¹Rudolf Arnheim, *Art and Visual Perception: A Psychology of the Creative Eye*, 1954, p. 449-454. [64]

Caso Studio: “Dreams” di Akira Kurosawa

Ricercando legami tra cinema ed espressionismo non si può non citare “Dreams”¹², uno degli ultimi film diretti da Akira Kurosawa, opera autobiografica nella quale il regista racconta otto suoi sogni.

Viene preso in analisi il quinto episodio, “Corvi”, che rappresenta un momento altamente simbolico all’interno del film. Akira Kurosawa, attraverso l’alter ego protagonista, non solo rende omaggio a Vincent Van Gogh, ma esplora anche il modo in cui l’arte influenza la percezione della realtà.

Sebbene Van Gogh appartenga al Post-Impressionismo e non all’Espressionismo, il suo uso espressivo del colore e la sua rappresentazione soggettiva del mondo lo avvicinano a certe sensibilità espressioniste, infatti lo si riconosce come un precursore dell’espressionismo stesso¹³. Kurosawa, attraverso la sequenza di “Corvi”, sembra adottare una prospettiva che riecheggia questa visione, trasformando i quadri in ambienti immersivi e carichi di emotività, enfatizzando l’idea di una realtà percepita interiormente piuttosto che oggettivamente descritta.

La scelta di far letteralmente “entrare” il protagonista nei dipinti di Van Gogh non è casuale: essa enfatizza il concetto che la pittura non è semplicemente una riproduzione visiva della realtà, ma una trasformazione soggettiva ed emotiva del mondo circostante.



Figura 3.8: Frame tratto da “Dreams” di Akira Kurosawa.

L’Espressionismo si basa proprio su questa idea: non mostrare la realtà così com’è, ma come viene percepita interiormente dall’artista. Van Gogh, con il suo uso intenso e simbolico del colore e la pennellata vorticoso e dinamica, incarna perfettamente questa filosofia, e Kurosawa sfrutta la potenza visiva dei suoi quadri per immergere lo spettatore in un’esperienza onirica e viscerale.

Il passaggio dalla visione statica dei dipinti in un museo alla loro “esplorazione dall’interno” rappresenta il superamento della barriera tra spettatore e opera d’arte, che diventa un mondo a sé stante, vivo e pulsante.

¹²Akira Kurosawa, *Dreams*, 1990. [65]

¹³Norbert Wolf, *Expressionism (Taschen Basic Art)*, 2004, p. 14. [66]

3.4.4 Confronto tra i Modelli Scelti

A seguito della valutazione dei dati riscontrati sui modelli nel paragrafo 2.3, i migliori tre modelli per la generazione di immagini, tenendo conto di qualità, tempo di generazione medio e prezzo, risultano essere:

- Flux1.1 (Pro).
- Recraft V3.
- Imagen 3.

Questi tre modelli, con l'aggiunta di Midjourney che per motivi di API non era incluso nel benchmark analizzato, saranno testati e confrontati nel seguente paragrafo.

Per poter confrontare queste diverse IA, si chiederà a ciascuna di generare quattro diverse immagini, ognuna delle quali tratta dallo script generato nella sezione 3.3.3. Per ognuna immagine si definisce ora un prompt testuale da utilizzare come input:

1. Expressionist style. A ship with a yellow sail sails on a deep blue sea. The waves move slowly.
2. Expressionist style. Colorful hot air balloons (blue, green, and lilac) rise into the sunset sky.
3. Expressionist style. A blue desert under the moonlight. A caravan of dreamlike figures moves forward.
4. Expressionist style. A small blue house shaped like a piano rests quietly in a green-lilac meadow beneath a dreamy, starlit sky.

Nella generazione delle immagini, a differenza di quanto avviene con i modelli utilizzati per la scrittura degli script, l'utilizzo della lingua inglese sembra produrre risultati migliori rispetto all'italiano.

Questo è probabilmente dovuto al fatto che gli LLM, utilizzati per generare testo, sono allenati anche su task di traduzione, risultando quindi più capaci di interpretare correttamente prompt multilingua. Al contrario, i modelli di generazione di immagini generalmente non includono un allenamento specifico sulla traduzione del prompt, e ciò potrebbe spiegare la minore efficacia nell'elaborare descrizioni in lingue diverse dall'inglese.

Di seguito riporto un confronto di esempio con due immagini generate da Recraft V3:



Figura 3.9: Prompt in italiano.



Figura 3.10: Prompt in inglese.

Sebbene alcuni dei modelli presi in esame dispongano di funzionalità avanzate, come ad esempio l'inpainting, che consente la rigenerazione selettiva di parti specifiche di un'immagine, queste caratteristiche non verranno prese in considerazione nell'analisi seguente, poiché

l'obiettivo principale è confrontare le capacità di ciascun modello nella generazione visiva a partire dallo stesso prompt.

Flux1.1 (Pro)

Flux 1.1 (Pro)¹⁴ è un modello di intelligenza artificiale sviluppato da Black Forest Labs, basato su un'architettura ibrida che combina blocchi di trasformatori multimodali e modelli di diffusione parallela. È stato progettato per interpretare con precisione descrizioni testuali e generare immagini dettagliate e realistiche in breve tempo.

Il modello è accessibile attraverso diversi piani: Basic (11,90 dollari al mese), Pro (15,90 dollari al mese), Max (18,90 dollari al mese) e Pro Max (29,90 dollari al mese). Flux 1.1 (Pro) consente una prova gratuita che offre 10 crediti iniziali per la generazione di immagini. Dispone, inoltre, di alcune funzionalità avanzate, come la modalità Ultra (generazione di immagini ad alta risoluzione fino a 4 megapixel) e la modalità Raw (immagini fotografiche iper-realistiche), che sono accessibili a tutti gli utenti abbonati.



Figura 3.11: Prima immagine generata da Flux1.1 (Pro).



Figura 3.12: Seconda immagine generata da Flux1.1 (Pro).



Figura 3.13: Terza immagine generata da Flux1.1 (Pro).



Figura 3.14: Quarta immagine generata da Flux1.1 (Pro).

Recraft V3

Recraft V3¹⁵ è un modello di intelligenza artificiale sviluppato da Recraft, specializzato nella generazione di immagini da descrizioni testuali e progettato specificamente per grafica vettoriale e raster. Al suo interno integra capacità avanzate di interpretazione dei prompt, consentendo agli utenti un controllo dettagliato sul posizionamento di testo ed elementi grafici nelle immagini generate. Include, inoltre, strumenti di editing avanzati per le immagini, come l'inpainting e l'outpainting, che consentono di modificare e ampliare le immagini generate.

Il modello è accessibile attraverso diversi piani: Free (50 crediti giornalieri), Basic (costo di 12 dollari al mese), Advanced (costo di 24 dollari al mese) e Pro (costo di 48 dollari al mese).

¹⁴Black Forest Labs, *FLUX.1 Documentation*, 2024. [67]

¹⁵Recraft AI, *Recraft Documentation*, 2024. [68]



Figura 3.15: Prima immagine generata da Recraft V3.



Figura 3.16: Seconda immagine generata da Recraft V3.



Figura 3.17: Terza immagine generata da Recraft V3.



Figura 3.18: Quarta immagine generata da Recraft V3.

Imagen 3

Imagen 3¹⁶ è un modello di intelligenza artificiale sviluppato da Google DeepMind, basato su modelli di diffusione latente, specializzato nella generazione di immagini dettagliate e altamente realistiche a partire da prompt testuali. Al suo interno integra capacità avanzate di interpretazione del linguaggio naturale, offrendo agli utenti immagini precise e in una vasta gamma di stili, dal fotorealismo a rappresentazioni artistiche astratte.

Il modello è attualmente accessibile gratuitamente attraverso piattaforme come ImageFX e Gemini AI.



Figura 3.19: Prima immagine generata da Imagen 3.



Figura 3.20: Seconda immagine generata da Imagen 3.



Figura 3.21: Terza immagine generata da Imagen 3.



Figura 3.22: Quarta immagine generata da Imagen 3.

¹⁶Google Cloud Platform, *Imagen3 Documentation*, 2024. [69]

Midjourney

Midjourney¹⁷ è un modello di intelligenza artificiale sviluppato da Midjourney, Inc., basato su reti neurali e modelli di diffusione, specializzato nella generazione di immagini a partire da descrizioni testuali. Al suo interno integra, quindi, sia capacità avanzate di interpretazione dei prompt testuali che funzionalità innovative come la modifica o estensione di immagini già generate (outpainting).

Il modello è accessibile esclusivamente tramite abbonamenti a pagamento, poiché la prova gratuita è stata sospesa da marzo 2023. Sono disponibili diversi piani che differiscono principalmente per il tempo di calcolo mensile assegnato (espresso in minuti di consumo delle GPU) e l'accesso a funzionalità aggiuntive, come la generazione prioritaria delle immagini.



Figura 3.23: Prima immagine generata da Midjourney.



Figura 3.24: Seconda immagine generata da Midjourney.



Figura 3.25: Terza immagine generata da Midjourney.



Figura 3.26: Quarta immagine generata da Midjourney.

3.4.5 Confronto Output e Scelta del Modello

A seguito dell'analisi approfondita dei modelli di generazione di immagini presentati nei paragrafi precedenti, si procede ora con un confronto sistematico dei risultati ottenuti da ciascun modello IA testato.

Il fine di questa valutazione è identificare la soluzione più adeguata alla realizzazione del video finale, garantendo un equilibrio tra qualità estetica, coerenza stilistica, fedeltà alle emozioni da trasmettere e praticabilità operativa.

Per effettuare un confronto equo tra i modelli analizzati, sono stati adottati i seguenti criteri:

- Qualità dell'immagine: Valutazione della definizione dell'immagine, presenza di artefatti e fedeltà alla richiesta.
- Adattabilità alla corrente artistica scelta: Analisi della capacità del modello di generare immagini in linea con il concept visivo definito.

¹⁷Midjourney, *Midjourney Documentation*, 2024. [70]

- **Coerenza narrativa:** Coerenza tra immagini generate in sequenza per garantire una narrazione fluida nel video.

Per quanto riguarda l'adattabilità alla corrente artistica scelta, in questo caso l'Espressionismo, si sono valutati quelli che sono i caratteri principali¹⁸ di questa corrente, ovvero:

- **Uso espressivo del colore,** che è generalmente saturo e non naturalistico, utilizzato per evocare stati d'animo intensi piuttosto che per riprodurre fedelmente la realtà.
- **Distorsione delle forme,** che mira a enfatizzare la soggettività della percezione piuttosto che la fedeltà alla realtà.
- **Impatto emotivo,** per trasmettere emozioni in modo diretto e viscerale, utilizzando immagini fortemente evocative.

Riprendendo il metodo di analisi che è stato svolto nel paragrafo 2.2.2 per la valutazione delle IA nel task di scrittura creativa (in quel caso operata da Lech Mazur¹⁹), questa comparazione si è svolta tramite l'utilizzo di ChatGPT 4o:

Definiti i parametri di valutazione, son state date in input tutte le immagini generate e si è chiesto di assegnare un punteggio da uno a dieci a ciascuna IA per ogni criterio di valutazione precedentemente descritto.

Di seguito sono riportati i risultati di questa comparazione:

IA di Image Generation	Qualità dell'Immagine	Adattabilità alla Corrente Artistica Scelta	Coerenza Narrativa
Flux1.1 (Pro)	6	5	6
Recraft V3	8	9	7
Imagen 3	7	6	6
Midjourney	7	7	7

Tabella 3.3: Confronto tra le IA di image generation.

¹⁸Norbert Wolf, *Expressionism (Taschen Basic Art)*, 2004, p. 14-31. [66]

¹⁹Lech Mazur, *LLM Creative Story-Writing Benchmark*, 2025. [26]

Si riportano, inoltre, quelle che sono state le motivazioni che l'IA ha fornito a fronte di queste valutazioni:

Motivazioni delle valutazioni:

- Qualità dell'Immagine: Midjourney e Imagen 3 hanno prodotto immagini dettagliate e ben definite, con un punteggio leggermente superiore a Flux1.1 Pro. Tuttavia, Recraft V3 si distingue per una qualità visiva più espressiva, anche se non necessariamente più dettagliata in senso tecnico.
- Adattabilità alla Corrente Artistica Scelta: Recraft V3 si è dimostrato il modello più adatto all'Espressionismo, grazie al forte utilizzo del colore, alla distorsione delle forme e all'intensità emotiva. Flux1.1 Pro ha mostrato più difficoltà nel replicare le caratteristiche espressioniste rispetto agli altri modelli.
- Coerenza Narrativa: Midjourney ha mostrato una buona coerenza nella generazione di immagini sequenziali, mentre Flux1.1 Pro e Imagen 3 sono risultati meno consistenti. Recraft V3, pur eccellendo nello stile espressionista, ha avuto una variazione maggiore nella resa di immagini simili, ottenendo un punteggio intermedio.

Scelta quindi l'IA di image generation, si procederà nel prossimo paragrafo alla creazione dello storyboard per l'elaborato.

3.4.6 Generazione dello Storyboard tramite Recraft V3

Recraft V3 al suo interno dà la possibilità agli utenti di creare “lavagne” di lavoro, per poter visualizzare e organizzare le immagini mantenendo sempre una visione d'insieme.

Il processo di creazione dello storyboard ha seguito diversi fasi:

1. Creazione generazione di immagini: si imposta la tipologia di stile che si vuole ottenere, caratteristica fondamentale in quanto Recraft V3 implementa nativamente lo stile espressionista. Dopo di che, si imposta il formato dell'immagine, ovvero il 16:9, per adattarlo già al formato del video, ed infine si immette il prompt di descrizione proposto nello script tradotto in inglese. Questo è un processo iterativo, volto ad ottenere il risultato migliore possibile o il più facilmente lavorabile.
2. Se un'immagine non risulta in linea con le aspettative, in relazione al problema dell'immagine, si procede in maniere diverse:
 - Se l'immagine non è stata creata in maniera adeguata, si procede con una nuova generazione senza cambiare nulla. Qualora anche la nuova generazione non sia ritenuta soddisfacente, si va ad agire sul prompt, specificando meglio ciò che si vuole ottenere.
 - Se solo una porzione di immagine non risultasse adatta, si va ad utilizzare la funzione *Edit Area*, che permette di selezionare una porzione di immagine e immettere un nuovo prompt per specificare cosa si vuole andare ad inserire o modificare. Questo fa sì che solamente la porzione d'area selezionata venga modificata.

- Le immagini appartenenti alla stessa scena vengono generate sfruttando la funzione *Image Colors*, che permette generare un'immagine avendo come riferimento una palette di colori, in modo tale da poter dar continuità visiva e cromatica alla scena.

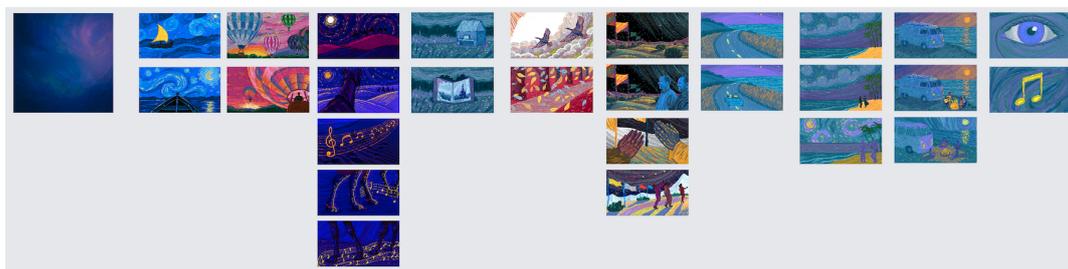


Figura 3.27: Lavagna di lavoro su Recraft V3.

Modifica di immagini con Photoshop

In alcuni casi, le immagini generate o modificate su Recraft V3 non risultavano adeguate per quanto era richiesto da script. In questi casi, si è ricorsi ad una modifica manuale su Photoshop²⁰, programma per modificare le immagini, per ottenere quindi il risultato voluto. In particolare, questo processo si è reso necessario per due casistiche in particolare:

- Per la seconda scena, quella delle mongolfiere, in quanto Recraft V3 non riusciva a generare un “uomo che suona il violoncello”. Di conseguenza, tramite Photoshop, si è rimosso dall'immagine l'uomo di spalle presente nell'immagine tramite lo strumento *Brush Tool*, che funziona in maniera simile al *Timbro Clone*, permettendo quindi di coprire parti di immagine con altre porzioni della stessa, come mostrato in figura 3.28. In seguito si è fatto generare appositamente un uomo che suona il violoncello da Recraft V3 e su Photoshop lo si è scontornato e importato nella prima immagine, come mostrato in figura 3.29.

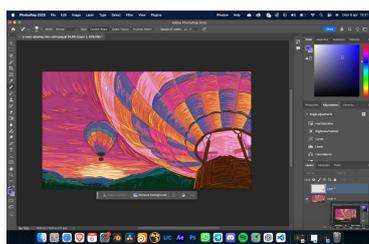


Figura 3.28: Rimozione dell'uomo di spalle.

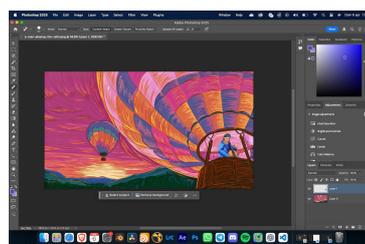


Figura 3.29: Inserimento dell'uomo che suona il violoncello.

- La scena, però, più complicata da realizzare è stata la settimana, ovvero: *Una figura guida una decappottabile blu, lasciandosi dietro note musicali*. La realizzazione è stata complessa in quanto si aveva il bisogno di visualizzare principalmente tre immagini per poter poi realizzare i video tramite la funzione *Pose-To-Pose*: un'immagine della strada costiera vuota, un'immagine con la macchina che entrasse in scena ed una in cui la macchina fosse più avanti nella strada lasciandosi dietro le note musicali. Si è quindi richiesto a Recraft V3 la generazione della strada vuota, e poi, copiata ed incollata quest'ultima, tramite lo strumento di *Edit Area*, si è aggiunta la decappottabile blu

²⁰Adobe Inc., *Adobe Photoshop*, 2025. [71]

(3.30). A questo punto si è resa evidente un'altra problematica dell'IA, ovvero la non replicabilità: è infatti impossibile, tramite lo stesso metodo, generare di nuovo la stessa macchina, ma spostata più avanti nella strada. Si è quindi proceduto con i seguenti passaggi:

1. Tramite Photoshop, si è ritagliata il profilo della decappottabile blu e coperto la parte della strada in cui si trovava con lo strumento *Brush Tool*. La strada mostra però una curva verso destra, quindi l'immagine ritagliata della decappottabile è stata specchiata, e manualmente si è invertita la posizione dei due passeggeri, come mostrato in figura 3.31.
2. Per aggiungere le note musicali, si è esportata l'immagine precedente e la si è caricata nuovamente su Recraft V3, dove, con lo strumento *Edit Area* si sono generate le note musicali, come mostrato in figura 3.32.
3. Queste, però, non erano sufficienti per coprire l'intera strada, quindi, tramite nuovamente Photoshop, si sono scontornate e duplicate una serie di note musicali, ingrandendole verso il margine inferiore dell'immagine per far sembrare che stessero "volando" (figura 3.33).



Figura 3.30: Immagine della strada costiera con la decappottabile blu generata da Recraft V3.



Figura 3.31: Immagine della strada costiera dopo la modifica con Photoshop.



Figura 3.32: Aggiunta delle note con Recraft V3.



Figura 3.33: Duplicazione delle note con Photoshop.

In totale, per realizzare lo storyboard, sono state generate circa 350 immagini, in quanto Recraft V3 fornisce due output possibili ogni volta che si effettua una generazione, di cui poi 28 sono andate a comporre lo storyboard. Il basso tasso di ottenimento dell'immagine generata evidenzia quindi la non immediatezza nell'ottenere ciò che si desidera, seppur lo strumento lo permetta.

A partire da queste immagini, si procederà con la generazione dei video.

Stereotipizzazioni nelle immagini generate

Prima di procedere con la generazione dei video, è però doveroso effettuare un'ulteriore considerazione. Durante la generazione dello storyboard, è emerso un comportamento si-

gnificativo che illustra chiaramente un pattern di stereotipizzazione algoritmica: quando ad un'immagine rappresentante un "deserto di notte" si volevano aggiungere dei volti di donne tramite lo strumento *Edit Area*, il sistema produceva sistematicamente rappresentazioni di donne con velo, indipendentemente da specifiche aggiuntive, come mostrato nelle figure 3.34 e 3.35. Questo fenomeno non si limitava a una singola generazione, ma si ripeteva in modo costante, evidenziando la presenza di un bias culturale nel modello.



Figura 3.34: Prima prova di generazione.



Figura 3.35: Seconda prova di generazione.

Questo comportamento rivela come i modelli di IA generativa incorporino e riproducano associazioni stereotipate presenti nei dataset di addestramento: in questo caso il sistema probabilmente opera attraverso un automatismo che collega ambiente desertico, cultura araba o islamica e rappresentazione femminile velata, creando una catena associativa rigida e deterministica.

Ciò evidenzia una problematica significativa: questo automatismo algoritmico limita la rappresentazione della diversità culturale e perpetua una visione monolitica delle identità femminili in determinati contesti geografici.

L'identificazione di questo pattern sottolinea l'importanza di un approccio critico nell'utilizzo dell'IA generativa. La ricerca richiede una consapevolezza costante dei bias impliciti nei modelli, lo sviluppo di strategie di prompt engineering più inclusive e una validazione continua dei contenuti generati rispetto agli obiettivi di rappresentazione equa.

L'osservazione documenta, inoltre, come la neutralità apparente della tecnologia nasconda in realtà meccanismi di perpetuazione di stereotipi che necessitano di essere identificati, analizzati e contrastati attraverso pratiche consapevoli di utilizzo degli strumenti di IA generativa.

3.5 Generazione dei Video

Dopo aver definito lo storyboard visivo e aver strutturato le sequenze narrative, si è passati alla fase di realizzazione video, utilizzando strumenti di generazione automatica basati su intelligenza artificiale. In questa fase, l'obiettivo principale è stato quello di tradurre i keyframe visivi e i prompt descrittivi precedentemente definiti in brevi clip video coerenti, fluidi e stilisticamente allineati alla visione creativa complessiva del progetto.

La scelta della piattaforma di generazione ha tenuto conto non solo della qualità dei risultati prodotti, ma anche della flessibilità nei controlli offerti, dei costi e della sostenibilità del processo produttivo nel tempo. Il sistema adottato, Adobe Firefly, ha consentito di implementare due modalità differenti di generazione (Image-to-Image Video e Image-to-Video), offrendo un buon compromesso tra controllo, qualità visiva e accessibilità.

In questo paragrafo verrà descritto il processo tecnico e pratico di generazione, analizzando le impostazioni adottate, le problematiche incontrate e le scelte progettuali.

3.5.1 Generazione dei Video Tramite Adobe Firefly

A seguito dell'analisi comparativa dei principali sistemi di generazione video mediante intelligenza artificiale (par. 2.4), si è deciso di utilizzare Adobe Firefly²¹ per la realizzazione del prodotto finale. La scelta è ricaduta su questo strumento per tre motivi principali:

- la presenza di un'interfaccia accessibile e completamente online (non richiede installazione o configurazione locale).
- la disponibilità di modalità di generazione differenziate (Image-to-Image Video e Image-to-Video) che si sono dimostrate particolarmente adatte alle esigenze del progetto.
- l'integrazione diretta di opzioni di configurazione come l'inquadratura, l'angolo di ripresa e i movimenti di camera.

Nonostante alcune criticità emerse in corso d'opera — legate ai costi e alla variabilità dei risultati — Adobe Firefly ha rappresentato una soluzione efficace per garantire coerenza stilistica, rapidità produttiva e flessibilità nel controllo dei contenuti generati.

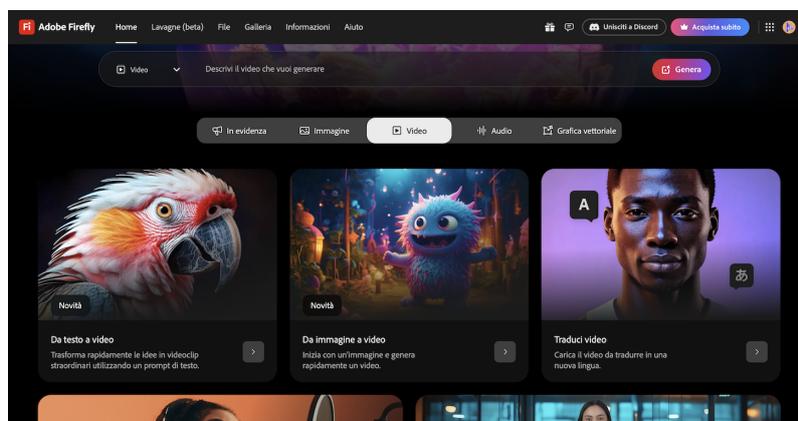


Figura 3.36: Interfaccia iniziale di Adobe Firefly.

²¹Adobe Inc., *Adobe Firefly*, 2025. [40]

La piattaforma online in questione offre diverse modalità di generazione video:

- Text-to-Video
- Image-to-Video
- Image-to-Image Video

La piattaforma integra inoltre strumenti per la generazione di immagini e contenuti audio, che tuttavia non sono stati impiegati nel presente progetto.

Adobe Firefly genera contenuti video della durata standard di cinque secondi, disponibili in diverse risoluzioni che comportano costi di generazione differenziati:

- Risoluzione 540p: 100 crediti per generazione
- Risoluzione 720p: 250 crediti per generazione
- Risoluzione 1080p: 500 crediti per generazione

Il processo di generazione video si è articolato su un periodo bimestrale, in considerazione dei limiti mensili di generazione previsti dal sistema. Mediante l'abbonamento "Adobe Firefly Pro", al costo di 33,82€ mensili, sono stati messi a disposizione 7.000 crediti generativi, corrispondenti alla produzione di settanta video mensili.

Il sistema di abbonamento prevede le seguenti opzioni tariffarie:

- "Adobe Firefly Free": si hanno 10 generazioni gratuite al mese.
- "Adobe Firefly Standard": 11,26€ mensili per 2.000 crediti generativi.
- "Adobe Firefly Premium": 225,53€ mensili per 50.000 crediti generativi.

Un elemento di criticità significativa, non adeguatamente considerato nella fase di pianificazione iniziale, è rappresentato dalla variabilità dei costi di queste risorse. Oltre all'elevato costo e alla limitata personalizzazione, la volatilità dei prezzi da un mese all'altro rende particolarmente complessa l'organizzazione di un piano di produzione strutturato. Considerando la durata prevista del video secondo lo script elaborato, il fabbisogno di generazioni sarebbe stato:

$$\frac{\text{Durata Video in Secondi}}{\text{Durata della Singola Generazione}} = \text{Numero Generazioni Totali}$$

Nel caso specifico:

$$\frac{300s}{5s} = 60$$

Conseguentemente, la produzione dell'intero video avrebbe richiesto approssimativamente un mese completo di generazioni. Tale stima non tiene conto della possibilità di errori o di generazioni non soddisfacenti, risultando pertanto sottodimensionata rispetto al fabbisogno effettivo. Successivamente alla generazione dei primi settanta video nel corso del primo mese, Adobe ha quintuplicato i costi di generazione per i video in risoluzione 1080p (precedentemente fissati a 100 crediti), rendendo necessario il ricorso alla generazione di un maggior numero di video in risoluzione 540p, in considerazione del materiale ancora da

produrre per il completamento del progetto. Tale circostanza imprevista avrebbe potuto compromettere la realizzazione del progetto. Tuttavia, la decisione adottata durante il primo mese di ridurre al 50% la velocità dei video in fase di post-produzione (aspetto che sarà approfondito nella sezione dedicata al montaggio) ha consentito di dimezzare il numero di generazioni richieste, permettendo il completamento del video senza ulteriori criticità.

Adobe Firefly consente inoltre la configurazione di diverse caratteristiche del video:

- Modello di intelligenza artificiale: viene selezionato automaticamente il modello *Firefly Video*, con la possibilità di utilizzare alternativamente i modelli *Veo2* e *Veo3*
- Proporzioni del video: possibilità di selezione tra i formati 16:9, 9:16 o 1:1
- Tipologia di inquadratura: definizione del tipo di shot desiderato (Extreme Wide Shot, Wide Shot, Medium Shot, Close Up, Extreme Close Up)
- Angolo di ripresa: selezione tra ripresa aerea, a livello degli occhi, dall'alto, dal basso, dall'alto verso il basso
- Movimento di camera: impostazione della tipologia di movimento (Zoom In, Zoom Out, Carrellata verso destra, Carrellata verso sinistra, Tilt, Pan)

Il sistema offre inoltre la possibilità di definire le immagini di inizio e fine del video, elemento fondamentale per l'implementazione della tecnica *Image-to-Image Video*. È opportuno precisare che non è necessario impostare entrambe le immagini, potendo definire esclusivamente quella di partenza per ottenere la tecnica *Image-to-Video*. Infine, è disponibile uno spazio dedicato alla definizione di un prompt testuale. Nel caso dell'utilizzo della tecnica *Image-to-Image Video*, le impostazioni relative alla tipologia di inquadratura, al tipo di shot e al movimento di camera risultano non utilizzabili.

Conformemente al workflow descritto nel paragrafo 3.1.1, sono state generate sostanzialmente due tipologie di video:

- Video generati mediante la tecnica *Image-to-Image Video*, utilizzata specificatamente per la realizzazione di transizioni e quando erano necessari passaggi precisi all'interno di una sequenza (come nel caso della scena con l'automobile sulla strada costiera, già analizzata nel paragrafo 3.4.6 relativamente alle modifiche effettuate con Photoshop).

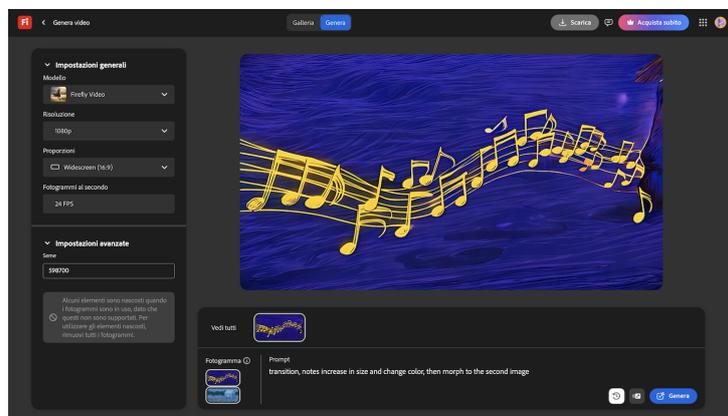


Figura 3.37: Esempio di generazione di una transizione.

Nel caso della figura 3.37, viene illustrato il processo di realizzazione di una transizione: oltre all'impostazione del primo e dell'ultimo fotogramma, mediante il prompt

testuale sono state ulteriori specifiche per il controllo delle modalità di evoluzione del cambiamento.

- Video generati mediante la tecnica *Image-to-Video*, impiegata quando non era necessario definire un fotogramma finale per il video in questione.

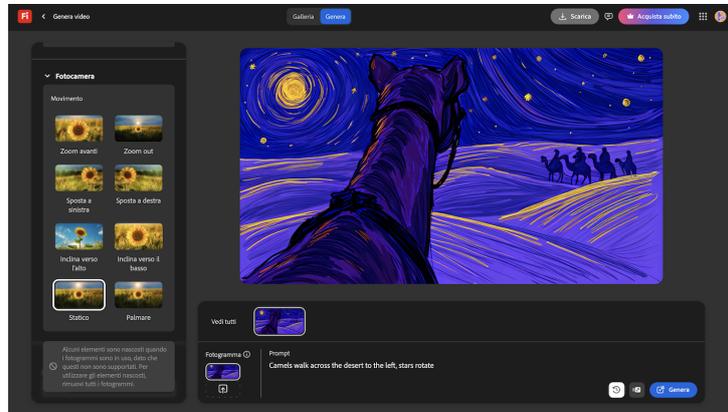


Figura 3.38: Esempio di generazione di un video tramite Image-to-Video.

La figura 3.38 illustra un esempio di questa tipologia di generazione secondo la configurazione adottata durante la realizzazione del video. Nello specifico, è stato selezionato il primo fotogramma, impostato un prompt testuale per descrivere gli eventi da sviluppare all'interno del video, e infine configurato il movimento di camera (in questo caso una ripresa statica).

Per la realizzazione del video finale sono stati generati complessivamente novantanove video. Considerando che nel montaggio finale sono presenti trentacinque video, si può stimare che, in media, per ogni generazione soddisfacente sono stati necessari circa tre tentativi. I video che hanno richiesto un maggior numero di generazioni sono stati principalmente le transizioni e le sequenze contenenti figure umane: in queste ultime, frequentemente il movimento risultava impreciso, con fenomeni di sovrapposizione delle parti corporee dei soggetti presenti.

Distorsioni dovute al processo di addestramento

Come già riscontrato precedentemente nell'analisi dello storyboard, anche durante la fase di generazione dei contenuti video si sono verificati diversi casi di bias culturali presenti nei video generati. Tale fenomeno è stato osservato esclusivamente nelle transizioni, ovvero nei passaggi in cui l'intelligenza artificiale disponeva di maggiore libertà creativa.

In particolare, si evidenziano due casi specifici:

- Durante la transizione tra il video del deserto e quello della strada costiera, si manifestavano bandiere di paesi europei, come evidenziato nella figura 3.39. Specificatamente, risultava particolarmente evidente la bandiera dell'Ucraina, mentre sullo sfondo era visibile quella dell'Austria.
- Durante la medesima transizione, veniva rappresentato un gruppo di uomini d'affari, come mostrato nella figura 3.40.



Figura 3.39: Bandiere di paesi europei.



Figura 3.40: Uomini d'affari.

Nonostante ripetute generazioni e tentativi di modifica del prompt testuale, questi due elementi continuavano a manifestarsi persistentemente.

La risoluzione di tale problematica è stata tuttavia fornita da Adobe Firefly stesso: tutte le prove relative a questa specifica transizione sono state realizzate durante il primo mese di generazione. Nel corso del secondo mese, invece, è stata sufficiente un'unica generazione, mantenendo invariato il prompt, per ottenere un risultato privo di entrambe le casistiche problematiche precedentemente riscontrate.

Pertanto, da questo cambiamento, si può dedurre che, probabilmente, sia stato aggiornato il modello di intelligenza artificiale, eliminando alcuni di questi bias derivanti dalla fase di addestramento.

3.6 Generazione Musica, Suoni e Voiceover

La componente sonora rappresenta un elemento centrale nel progetto audiovisivo, non solo come accompagnamento alle immagini, ma come strumento narrativo e affettivo capace di amplificare, evocare e strutturare l'esperienza emozionale dell'utente. In un contesto in cui tutti i contenuti visivi sono generati artificialmente e privi di audio nativo, si è reso necessario progettare e generare ex novo l'intera colonna sonora: musiche, effetti diegetici e voiceover.

In questo paragrafo si illustrerà il processo seguito per la creazione di questi elementi, evidenziando le scelte tecniche e creative adottate. Verranno analizzate le piattaforme di IA generativa utilizzate (Suno per la musica ed ElevenLabs per suoni e voce narrante), il modo in cui i prompt e i parametri sono stati affinati durante le prove.

3.6.1 Generazione della musica

All'interno del contesto per il quale il video sarebbe dovuto essere realizzato, la musica rivestiva un ruolo di fondamentale importanza, in quanto veicolo utilissimo per la trasmissione di emozioni. Inoltre, durante la fase di scrittura dello script, Claude 3.7 Sonnet ha delineato le linee guida anche per quella che sarebbe dovuta essere la base musicale, nello specifico:

- Canzone per il primo atto: una musica eterea, melodie sospese, realizzata tramite strumenti acustici, che potessero trasmettere gioia, nostalgia, malinconia, pace e relax.
- Canzone per il secondo atto: una musica con ritmi definiti, dinamici ed energici, che trasmetta vitalità, energia, libertà, calore e condivisione.

Quindi, si dovranno realizzare due canzoni per il video, ciascuna con i propri ritmi e le proprie finalità.

Per far ciò, si è utilizzato *Suno*²², strumento online di intelligenza artificiale generativa dedicato alla creazione musicale. Anche Suno, come gli altri strumenti online visti fin'ora, offre diversi piani di abbonamento:

- Free Plan: fornisce dieci generazioni gratuite al giorno, a discapito del modello utilizzabile, che è solamente il v3.5 (terz'ultima versione, in quanto sono disponibili anche la 4.0 e la 4.5). Inoltre, le canzoni generate non possono essere usate a scopi commerciali.
- Pro Plan: al costo di 8\$ al mese, fornisce la possibilità di generare fino a 500 canzoni al mese utilizzando il modello v4.5. Inoltre, fornisce la possibilità dell'utilizzo delle canzoni generate per scopi commerciali, ma solo fin quando è sottoscritto l'abbonamento. Infine, aggiunge funzioni per rendere più personalizzabile l'output, come la funzione "split song", che separa la canzone generata in diversi livelli, suddividendo i canali vocali da quelli strumentali.
- Premier Plan: al costo di 24\$ al mese, fornisce le stesse funzioni del Pro Plan, aumentando solamente il numero di generazioni mensili che, da 500, passa a 2000.

Ai fini realizzativi del video, è stata sufficiente la versione Free.

²²Suno Inc., *Suno*, 2023. [50]

A livello di funzionalità, Suno permette la generazione di basi sonore a partire da un prompt testuale, e fornisce la possibilità di specificare la tipologia di stile musicale che si vuole ottenere. È anche possibile specificare se si vuole una canzone solo musicale o con un testo cantato, e, se necessario, scrivere il testo specifico della canzone. Inoltre, si può inserire un audio o una base di riferimento da utilizzare come esempio, funzione non utilizzata per la generazione delle due basi per questo progetto. Infine, è importante evidenziare come Suno, per ogni prompt immesso, crea due output, al fine di poter fornire all'utente due opzioni tra cui scegliere.

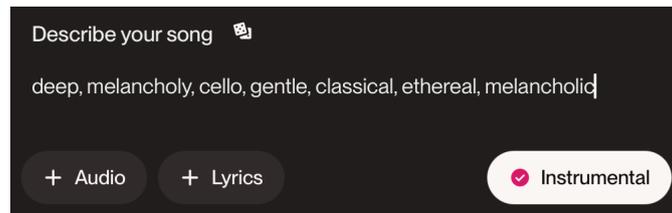


Figura 3.41: Pannello di generazione di Suno.

Nella figura 3.41 è mostrato il pannello di generazione di Suno. In particolare, è mostrato il prompt utilizzato per la generazione della prima canzone. Durante le diverse prove eseguite al fine di ottenere un risultato accurato, si sono infatti delineate alcune linee guida:

- Anche in questo caso, come per la generazione di immagini e di video, è preferibile utilizzare la lingua inglese per la scrittura del prompt.
- È risultato più utile fare un elenco di parole chiave rispetto ad una descrizione puntale della canzone, permettendo quindi più “libertà” allo strumento.

Per la generazione delle due canzoni sono state necessarie solamente 5 generazioni, con una percentuale di successo del 40%, a testimonianza di quanto sia efficiente questa piattaforma.

3.6.2 Generazione dei Suoni e del Voiceover

Per la generazione dei suoni e della voce narrante si è deciso di utilizzare *ElevenLabs*²³, uno dei più diffusi sistemi di intelligenza artificiale generativa per quanto riguarda il sonoro.

ElevenLabs offre molti piani di abbonamento:

- Piano Free: 10000 crediti mensili gratuiti, corrispondenti a circa 10 minuti di generazione vocale ad alta qualità. Sono incluse le funzioni di *text-to-speech*, *speech-to-text*, *Conversational AI*, il doppiaggio automatico e l'accesso alle API.
- Piano Starter: al costo di 5\$ mensili fornisce le stesse funzionalità del piano base, aumentandone i crediti fino a 30000, aggiungendo la funzione di *voice cloning* e dando la possibilità dell'utilizzo dei contenuti a fini commerciali.
- Piano Creator: al costo di 22\$ al mese fornisce le stesse funzionalità del piano Starter, ma incrementa i crediti (100000 crediti al mese) e la qualità dell'audio generato (fino a 192 kbps).

²³ElevenLabs Inc., *ElevenLabs*, 2023. [53]

- Piano Pro: al costo di 99\$ al mese fornisce le stesse funzionalità del piano Creator, aumentandone i crediti fino a 500000 e dando la possibilità di generare audio in 44.1 kHz PCM tramite le API.

Ai fini realizzativi del progetto è stato sufficiente l'utilizzo del piano free, in quanto non sarebbero state presenti tante linee di dialogo e i suoni utilizzati all'interno del video spesso si sarebbero ripetuti.

Generazione dei Suoni

I video generati tramite Adobe Firefly, come i video generati tramite altri modelli, sono sprovvisti di sonoro (ad eccezione di *Veo3* di Google, che, al tempo della realizzazione del video, non era ancora stato reso disponibile) e di conseguenza si è dovuto fare una sorta di “spoglio della sceneggiatura” specifico per gli effetti sonori.

Una volta identificati tutti i suoni che sarebbero serviti, si è utilizzato ElevenLabs per generare tutto il materiale. ElevenLabs permette questo tipo di funzione fornendo un prompt testuale per descrivere il suono, il parametro di durata per definirne la lunghezza e un'opzione per specificare quanto il modello debba essere fedele al prompt immesso (più il valore è alto e più il modello rispetterà quanto scritto, perdendo però “l'inventiva”).



Figura 3.42: Pannello di generazione dei suoni di ElevenLabs.

Nella figura 3.42 si mostra la generazione di un suono di “scintille”. Anche in questo caso, a seguito di diverse prove, è risultato che l'utilizzo della lingua inglese corrispondesse ad un migliore output.

Ogni richiesta ha un costo di 400 crediti e restituisce 5 output differenti, sufficientemente vari tra loro, offrendo così all'utente un'ampia gamma di scelta e riducendo la necessità di ripetere più volte il processo. Questo approccio consente quindi di ottenere più rapidamente il risultato desiderato. Per la generazione dei 19 suoni necessari, sono infatti stati effettuati circa 30 tentativi in totale, corrispondenti a una percentuale di successo pari a circa il 66%.

Generazione Vocale

Dallo script generato nel paragrafo 3.3.3 risulta come ogni scena presente nel cortometraggio dovesse essere accompagnata da una voce narrante di sottofondo. A seguito del benchmark valutato nel paragrafo 2.5.6, si è quindi deciso di procedere con la generazione vocale di questi testi tramite la funzionalità di *text-to-speech* fornita da ElevenLabs, in quanto offriva il miglior connubio tra qualità, efficienza e prezzo (che, come vedremo in seguito, non è risultato un parametro che inficiasse la scelta in quanto si è usufruito solamente delle generazioni gratuite mensili).

Nella sezione dedicata alla sintesi vocale, si possono modificare diversi parametri, avendo quindi ampio margine per la personalizzazione della tipologia di voce che si vuole ottenere:

- Voce da utilizzare: ElevenLabs mette a disposizione centinaia di voci tra cui poter scegliere, distinguendole tra voci maschili e femminili ed in base alla lingua del parlato. Provando diverse generazioni, si è constatato come le voci generate in lingua italiana suonassero molto “robotiche”, in quanto il bacino di dati da cui il modello attinge non è evidentemente molto vasto per l’italiano. Di conseguenza, si è deciso di generare le linee di dialogo interamente in inglese, e di aggiungere in fase di montaggio i sottotitoli in italiano.
- Modello per la generazione: nel caso specifico si è utilizzato l’*Eleven Multilingual v2*, in quanto la versione 3 include solamente un numero maggiore di lingue conosciute.
- Velocità del parlato.
- Stabilità: valori elevati di stabilità garantiscono una maggiore coerenza tra generazioni consecutive, ma possono al contempo rendere il risultato monotono.
- Similarità: incrementa la chiarezza della voce, ma se troppo alta può generare artefatti.
- Esagerazione dello stile: serve per modificare un audio fornito in input, quindi non è stato utile ai fini del progetto.
- Speaker Boost: aumenta ulteriormente la similarità a discapito dei tempi di generazione.

Il costo in crediti per la generazione vocale corrisponde ad un credito per ciascun carattere, spazi inclusi.

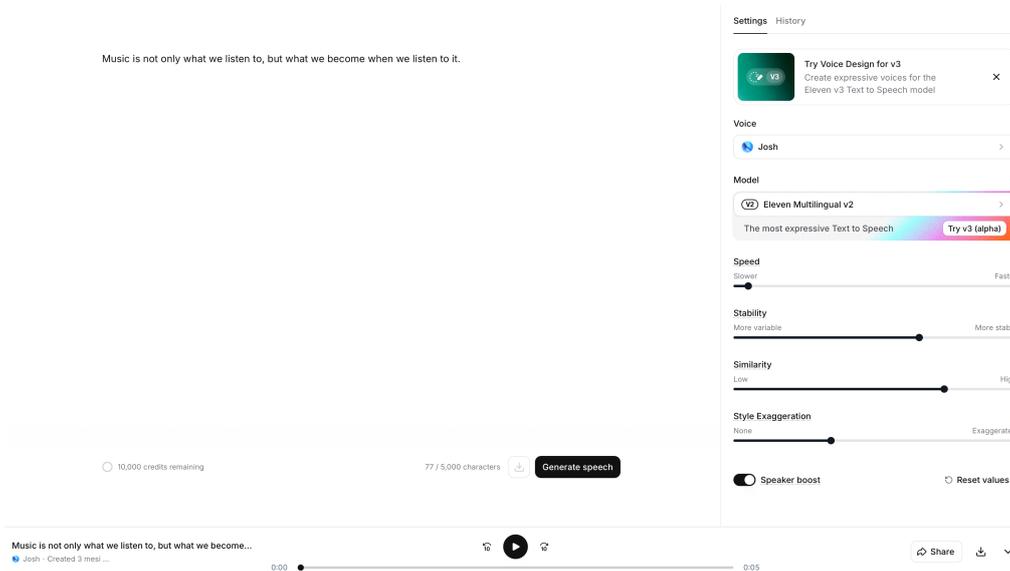


Figura 3.43: Pannello di generazione vocale di ElevenLabs.

Nella figura 3.43 è mostrato il modello utilizzato per il progetto. Prima di identificare la corretta combinazione di parametri sono state eseguite otto generazioni. Una volta trovati i valori corretti, però, non è più stato necessario ripetere alcuna generazione. Quindi per la generazione delle undici linee di dialogo presenti nel video sono state eseguite diciotto generazioni.

3.7 Montaggio del Video

Il montaggio finale è stato realizzato tramite il software *DaVinci Resolve*²⁴, cercando di seguire il più possibile le indicazioni fornite da Claude 3.7 Sonnet²⁵ nel momento di creazione dello script.

Un passaggio fondamentale è stato quello di organizzare fin da subito le clip in cartelle, al fine di mantenere il progetto il più ordinato possibile, considerato il numero di materiali a disposizione. Più precisamente, si è fatto uso di:

- 35 contenuti video.
- 19 contenuti sonori.
- 11 contenuti per il voiceover.
- 2 canzoni per la soundtrack.

Il progetto è stato poi inizializzato in Full HD, con un frame rate di 24 fps, esattamente come i video generati da Adobe Firefly²⁶.

Si è resa evidente fin da subito, però, un'ulteriore problematica che non era stata considerata: i video risultavano troppo fluidi rispetto alla scelta di rappresentazione visiva dell'espressionismo, causando una sorta di incongruenza tra quello che veniva mostrato e il come veniva mostrato.

Per risolvere questo problema si è deciso di rallentare tutti i video del 50%, facendo risultare in questo modo i video più “scattosi”, in quanto ora, al posto che avere 24 frame al secondo, se ne hanno 12. Questa scelta porta con sé un altro vantaggio: la durata massima dei video esportati da Adobe Firefly è passata dall'essere di cinque secondi a dieci, riducendo la necessità di ulteriori generazioni e, di conseguenza, i costi effettivi.

Di seguito è mostrata la timeline del progetto finale, con i vari livelli di tracce video, audio e dei sottotitoli.



Figura 3.44: Timeline del progetto su DaVinci Resolve.

²⁴Blackmagic Design, *DaVinci Resolve*, 2025. [72]

²⁵Anthropic, *Claude 3.7 Sonnet*, 2024. [61]

²⁶Adobe Inc., *Adobe Firefly*, 2025. [40]

Si è deciso, inoltre, di applicare un effetto di vignettatura generale, tipico del cinema. Per semplicità, siccome l'effetto era valido su tutta la timeline, si è deciso di applicarlo all'*Adjustment Clip* visibile nella traccia "Video 5" nella figura 3.44, in modo da non doverlo applicare ad ogni singola clip.

Effetto "Caleidoscopio" e transizione finale

La sfida più impegnativa a livello di montaggio è stata la realizzazione dell'effetto "caleidoscopio" finale, così come proposto dallo script elaborato da Claude 3.7 Sonnet. In particolare, l'indicazione data era la seguente: *Scene precedenti si sovrappongono, caleidoscopio di immagini*.

La problematica principale era legata al fatto che bisognasse, quindi, mostrare nove clip video da mostrare contemporaneamente e a cui si sarebbe andati ad applicare l'effetto "caleidoscopio". A seguito di alcune ricerche per provare a rendere automatico questo processo, si è reso necessario svolgerlo manualmente, in quanto i risultati ottenuti non erano soddisfacenti.

Si è quindi creata una *Fusion Composition*, in modo tale da poter lavorare in modo migliore e più controllato con le clip. Di seguito è mostrato l'albero di nodi utilizzato:

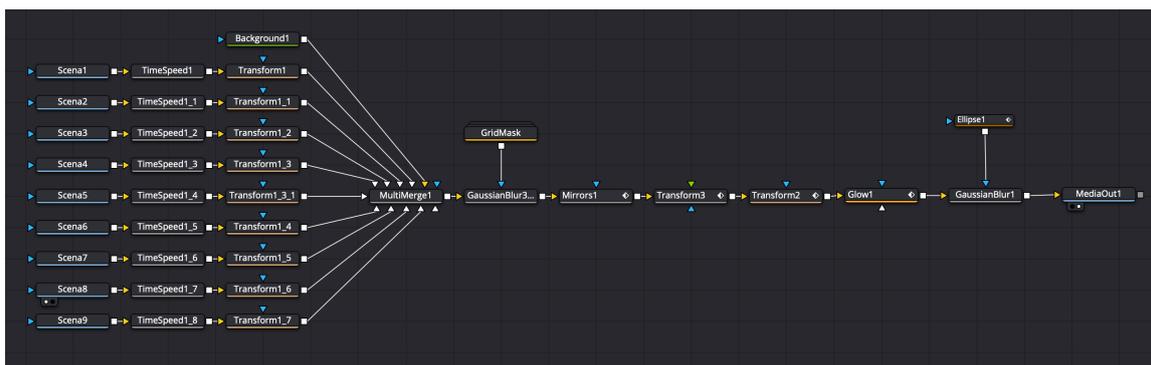


Figura 3.45: Albero dei nodi utilizzato per l'effetto "Caleidoscopio".

La logica alla base di questo albero di nodi è la seguente:

- Ogni clip era data come *MediaIn*, rallentata tramite il nodo *TimeSpeed* (per uniformare le clip a quelle mostrate nel cortometraggio) e disposta in una griglia 3x3 tramite il nodo *Transform*.
- Le nove clip disposte in griglia sono state poi combinate tramite il nodo *MultiMerge*.
- Si è poi applicato un effetto di *Gaussian Blur* mascherato sui bordi della griglia (l'insieme di nodi definito *GridMask*) per attenuare le discontinuità ai bordi delle clip.
- Sono stati animati il nodo *Mirrors* e i due nodi *Transform* per creare l'effetto di rotazione e movimento del caleidoscopio.
- Infine, si è applicato un effetto di *Glow* per distorcere leggermente i colori ed un ulteriore effetto di *Gaussian Blur* mascherato tramite un ellisse per andare a sfocare il contenuto verso i bordi dell'immagine.

Il risultato finale ottenuto è stato il seguente:



Figura 3.46: Caleidoscopio finale ottenuto.

Il passaggio successivo, come da script, è stato la transizione verso il “grande occhio blu che si apre e si chiude”. In questo caso, la tecnica utilizzata sinora per la generazione delle transizioni non ha prodotto risultati soddisfacenti, quindi anche in questo caso si è intervenuto manualmente, andando ad applicare una maschera sulla palpebra dell’occhio durante la chiusura, in modo che ricoprisse man mano il caleidoscopio, per poi riaprirsi a transizione avvenuta.



Figura 3.47: Transizione finale da caleidoscopio ad occhio.

seppur in numero minore, abbiano indicato emozioni quali “malinconia”, “nostalgia” e “relax”. Al contrario, il tema del viaggio è stato percepito con minore chiarezza: solo un partecipante lo ha esplicitamente menzionato.

2. “Hai percepito un cambiamento nel tuo stato d’animo dopo la visione del video? Se sì, in che modo?”:

Si è deciso di porre questa domanda per poter analizzare ulteriormente l’impatto del video a livello emotivo. Sebbene la maggior parte delle risposte sia stata negativa, una parte dei partecipanti ha fatto riferimento a sensazioni di rilassamento e tranquillità, confermando che tale emozione è effettivamente stata trasmessa.

3. “Se non avessi saputo come è stato realizzato il video, avresti detto fosse stato generato con intelligenza artificiale? 1 = Per nulla, 10 = Certamente”:

Dopo aver indagato le emozioni trasmesse, si è voluto capire quanto il video fosse stato percepito come realizzato tramite IA.

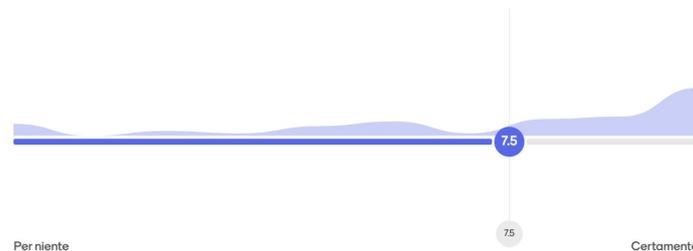


Figura 3.49: Slider di valutazione sulla percezione del video come generato da IA.

La valutazione finale corrisponde a 7.5, con una maggioranza di risposte che tendono verso il 10, votato da venti persone. Questo certifica che, nonostante lo stile artistico scelto, il video risulta comunque realizzato tramite IA. Per una valutazione corretta sul fatto che lo stile artistico abbia mitigato l’effetto di realizzazione tramite intelligenza artificiale sarebbe però servito un altro video con stile fotorealistico con cui poterlo paragonare.

4. “Con una parola, come descriveresti l’impatto complessivo del video?”:
- Per questa domanda si è voluto lasciare più spazio alle persone votanti, per vedere cosa effettivamente pensassero del video.



Figura 3.50: Nuvola di parole sull’impatto del video.

Come si può vedere nella nuvola di parole nella figura 3.50, le parole utilizzate sono state molte, con solamente sei parole che sono state scelte più di una volta, nello specifico “Colorato” (votato tre volte), “Banale” (votato due volte), “Caleidoscopico” (votato due volte), “Desolante” (votato due volte), “Pace” (votato due volte) e “Positivo” (votato due volte).

Questi dati offrono spunti interessanti: in primo luogo, l’impatto del colore è stato percepito correttamente, in linea con quanto previsto dall’analisi iniziale. Inoltre, è interessante la parola “Banale”, perché va a delineare un concetto che può essere oggetto di discussione: quanto i prodotti generati tramite intelligenza artificiale possano essere percepiti come innovativi o meno. In questo caso, la generazione dello script è probabilmente stata didascalica, senza aggiungere troppi livelli di introspezione e rimanendo spesso sul superficiale.

5. “Quanto, da 1 a 10, la scelta dell’uso di uno stile espressionista ha aiutato nel veicolare il messaggio? 1 = Per nulla, 10 = Completamente”:

Siccome la scelta di utilizzare uno stile espressionista è stata una delle decisioni più rilevanti all’interno di questo elaborato, si è voluto dedicare una domanda esplicitamente a questo.

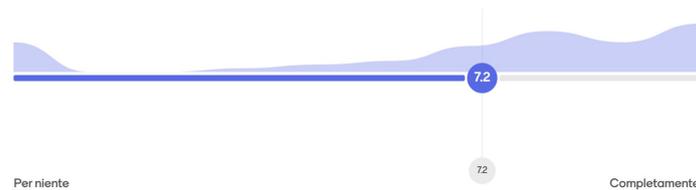


Figura 3.51: Slider di valutazione sulla scelta dello stile espressionista.

Nonostante la votazione complessiva presenti una media di 7,2, è importante considerare che, in assenza di un confronto con lo stesso video realizzato con stili differenti, manca un vero e proprio metro di paragone. Questa limitazione rende difficile trarre conclusioni definitive, come dimostra anche la distribuzione dei voti, fortemente polarizzata tra valutazioni molto alte e molto basse.

6. “Quanto, le distorsioni involontarie delle immagini hanno influenzato la visione del video? 1 = Per nulla, 10 = Completamente”:

Spesso nel video, alcuni passaggi subivano distorsioni involontarie, come ad esempio, il suonatore di violoncello, a cui scompariva lo strumento un frame dopo l’inizio del video. Siccome il tema delle distorsioni è stato anch’esso oggetto di ricerca all’interno di questo elaborato, si è voluto cercare di capire tramite questa domanda quanto queste imperfezioni abbiano inficiato la visione complessiva del video.

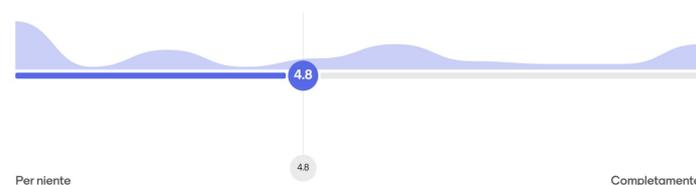


Figura 3.52: Slider di valutazione sulle distorsioni involontarie delle immagini.

La media dei voti, come mostrato nella figura 3.52, è di 4,8, risultato che evidenzia quanto queste distorsioni infastidissero, ma non rovinassero la visione complessiva.

Anche questa domanda, però, non fornisce una prova completa, in quanto bisognerebbe indagare in che misura le medesime distorsioni avrebbero inficiato la visione nel caso di un prodotto fotorealistico.

In conclusione, i dati raccolti suggeriscono che il video è riuscito in buona parte a trasmettere le emozioni desiderate, pur lasciando spazio a margini di miglioramento, specialmente per quanto riguarda la chiarezza del tema del viaggio e la profondità narrativa percepita. L'utilizzo di uno stile espressionista ha generalmente supportato l'esperienza emotiva, ma l'ambiguità visiva e alcune imperfezioni tecniche hanno probabilmente limitato un coinvolgimento più profondo.

Conclusioni

Il presente lavoro di ricerca si è posto l'obiettivo di investigare le potenzialità dell'intelligenza artificiale generativa nella produzione di contenuti audiovisivi, con particolare riferimento alla capacità di trasmettere emozioni analoghe a quelle provate durante un'esibizione musicale dal vivo. L'analisi dei risultati ottenuti, condotta attraverso il confronto tra i dati iniziali e i sondaggi conclusivi, ha evidenziato un bilancio complessivamente positivo dell'esperimento:

- Le emozioni percepite dal pubblico risultano sostanzialmente coerenti con quelle che si intendeva trasmettere;
- Il video ha generato un impatto emotivo positivo sul pubblico;
- La metodologia rappresentativa adottata si è dimostrata efficace nel veicolare il messaggio e facilitare la trasmissione emotiva.

Tuttavia, l'analisi dei risultati evidenzia anche alcune limitazioni significative. La principale criticità emersa riguarda la percezione di banalità del contenuto prodotto, aspetto che testimonia l'attuale incapacità dell'intelligenza artificiale di raggiungere livelli di creatività pienamente autonomi e innovativi nella narrazione.

Un ulteriore elemento di criticità è rappresentato dalla mancanza di coerenza stilistica complessiva del prodotto finale. Alcune sequenze appaiono infatti stilisticamente discordanti o concettualmente scollegate dal resto della narrazione. Tale problematica era tuttavia prevedibile data la natura del workflow adottato, e si può ragionevolmente ipotizzare che produzioni caratterizzate da budget più elevati e tempi adeguati per la personalizzazione dei dataset possano conseguire risultati qualitativamente superiori.

Questa è stata una tesi che ha cercato di spingere il più possibile l'intelligenza artificiale a "lavorare" da sola, cercando di apportare il minor numero di interventi possibili. Questo ha portato risultati interessanti, soprattutto mettendo in luce le problematiche per una produzione di questo tipo:

- Nella fase di analisi dei dati sono stati conseguiti risultati soddisfacenti, mentre le prime criticità sono emerse durante la composizione dello script. Il contenuto proposto si è rivelato eccessivamente didascalico, caratterizzato da una mancanza di introspezione e visione critica. È tuttavia opportuno sottolineare la precisione dimostrata dall'IA nella definizione dei timecode delle sequenze e nella costruzione di un ritmo narrativo coerente.
- La generazione delle immagini ha prodotto risultati interessanti, nonostante le limitazioni causate da un workflow basato su piattaforme online. Permangono tuttavia significative problematiche legate all'incoerenza stilistica e ai bias intrinseci nei modelli di intelligenza artificiale.

- La generazione video ha rappresentato la fase più problematica del processo, sebbene i risultati finali possano considerarsi accettabili. La mancanza di controllo preciso sui parametri di output costituisce una limitazione critica, rendendo i risultati sostanzialmente imprevedibili e richiedendo frequenti iterazioni per ottenere il prodotto desiderato. Anche in questo ambito sono emersi bias significativi.
- La generazione di contenuti audio musicali ha fornito risultati soddisfacenti, contrariamente a quanto osservato per la sintesi vocale. Nonostante la scelta strategica di utilizzare la lingua inglese per ampliare il dataset di riferimento, la qualità della voce generata è risultata poco convincente.

3.9 Sviluppi futuri

L'analisi condotta conferma come l'intelligenza artificiale sia ancora considerevolmente distante dalla possibilità di sostituire completamente il lavoro umano nel processo creativo, presentando limitazioni significative e problematiche non trascurabili. Emerge tuttavia chiaramente il potenziale di questa tecnologia come strumento di supporto, particolarmente efficace nell'ottimizzazione di specifici processi e nella generazione di idee.

È inoltre da considerare il carattere dinamico e in continua evoluzione di questa tecnologia. L'avanzamento registrato nell'ambito dell'intelligenza artificiale negli ultimi mesi potrebbe rendere alcune delle conclusioni qui presentate già parzialmente obsolete. Questo aspetto rappresenta simultaneamente un vantaggio e una sfida: da un lato offre accesso a strumenti progressivamente più performanti, dall'altro richiede un investimento temporale considerevole per rimanere aggiornati sugli sviluppi più recenti del settore.

Proprio questa rapidità di evoluzione della tecnologia rende difficile il poter fare stime sui tempi con cui questa tecnologia arriverà ad essere effettivamente utilizzabile in grandi produzioni: basti pensare che fino a qualche mese fa sarebbe stato impensabile immaginare un video con audio annesso generato con un solo script, e invece ora è stato rilasciato Veo3, che fa proprio questo.

Seppur l'innovazione procede con un grande ritmo, per la tipologia di workflow utilizzato in questa tesi non basterà il solo miglioramento stilistico o la potenza di calcolo: serviranno funzioni apposite, da introdurre sui siti delle varie intelligenze artificiali, che amplino il controllo umano su quello che viene generato.

L'introduzione di nuove funzionalità comporterà inevitabilmente un aumento dei costi. Già durante il periodo di elaborazione di questa tesi, molti servizi hanno incrementato le proprie tariffe oppure, laddove ciò non è avvenuto, hanno ridotto le specifiche offerte.

Così, uno strumento come l'intelligenza artificiale — inizialmente concepito come mezzo di democratizzazione dell'arte, in grado di permettere a chiunque di creare attraverso testo e immaginazione — rischia di diventare inaccessibile proprio a causa dei costi.

Contrariamente a queste tendenze, però, sembra che dal lato hardware ci si stia muovendo verso la creazione di dispositivi pensati su misura per far processare in locale i sistemi di intelligenza artificiale: NVIDIA è sul punto di lanciare i suoi mini computer che, essendo ottimizzati per queste precise funzioni, potrebbero abbattere i costi necessari. Ciò renderebbe tutti quei workflow che si basano su un processo locale decisamente più accessibili, richiedendo quindi solamente conoscenze tecniche specifiche per poter utilizzare questi sistemi.

Un punto fondamentale da trattare nei prossimi anni sarà la chiarificazione sul diritto d'autore riguardo l'utilizzo di queste tecnologie: al momento appare ancora poco chiara la legislazione a riguardo, rischiando quindi di poter generare contenziosi legali significativi tra creatori originali e utilizzatori di IA. L'ambiguità normativa attuale crea un vuoto giuridico che potrebbe scoraggiare l'investimento in progetti creativi basati su IA, limitando al contempo la tutela dei diritti degli artisti le cui opere vengono utilizzate nei dataset di addestramento.

Infine, sarà necessario definire con precisione i criteri di attribuzione della paternità artistica quando l'intervento umano si limita alla supervisione del processo automatizzato, stabilendo se e in che misura i contenuti generati dall'IA possano essere considerati opere derivate e quindi soggette alle relative normative sul copyright.

Ringraziamenti

Voglio dedicare quest'ultima parte della tesi a tutte le persone che, nel corso degli anni, mi sono state vicine e mi hanno aiutato.

Innanzitutto, grazie ai miei genitori, Andrea e Tiziana, e a mia sorella Sara, che mi hanno sempre sostenuto nei momenti di difficoltà. Un ringraziamento speciale va anche ai miei nonni, per tutte le candele accese prima di ogni esame.

Grazie di cuore a tutti i miei amici e amiche di Alessandria, che hanno saputo alleggerire i momenti di stress lungo questo percorso, e agli amici di Torino, con cui ho condiviso notti insonni piene d'ansia, ma anche tante gioie.

Infine voglio ringraziare i miei relatori, il professor Bottino e la professoressa Mazali, e Roberto e Claudia di Narrazioni Parallele, per avermi accompagnato e supportato nel percorso di tesi.

Bibliografia

- [1] Margaret A. Boden. *Computer Models of Creativity*. 2009. DOI: 10.1609/AIMAG.V30I3.2254. URL: <https://doi.org/10.1609/aimag.v30i3.2254>.
- [2] J. Walter Thompson Amsterdam, ING e Microsoft. *The Next Rembrandt: Recreating the work of a master with AI*. 2016. URL: <https://news.microsoft.com/europe/features/next-rembrandt/>.
- [3] Refik Anadol. *Machine Hallucinations: Nature Dreams*. 2021. URL: <https://refikanadol.com/works/machine-hallucinations-nature-dreams/>.
- [4] Ahmed Elgammal et al. *CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms*. 2017. arXiv: 1706.07068 [cs.AI]. URL: <https://arxiv.org/abs/1706.07068>.
- [5] Eva Cetinic e James She. *Understanding and Creating Art with AI: Review and Outlook*. 2021. arXiv: 2102.09109. URL: <https://arxiv.org/abs/2102.09109>.
- [6] Art Recognition. *Art Recognition: AI-based Art Authentication*. 2021. URL: <https://www.art-recognition.com/>.
- [7] Mario Klingemann. *Memories of Passersby I*. 2018. URL: <https://www.sothebys.com/en/auctions/ecatalogue/2019/contemporary-art-day-auction-119021/lot.109.html>.
- [8] Aaron Hertzmann. *Visual Indeterminacy in Generative Neural Art*. 2019. arXiv: 1910.04639. URL: <http://arxiv.org/abs/1910.04639>.
- [9] Kate Vass. *Helena Sarin: Neural Bricolage*. 2024. URL: <https://www.katevassgalerie.com/blog/helena-sarin-neural-bricolage>.
- [10] Christie's. *Is artificial intelligence set to become art's next medium?* 12 Dec. 2018. 2018.
- [11] Davide Tromba. *AI Storytelling and Cinematography: come i modelli open source stanno rivoluzionando la pipeline cinematografica*. Torino, Italia, 2025.
- [12] Peter Luisi. *The Last Screenwriter*. 2024. URL: <https://lastscreenwriter.com>.
- [13] Pei Sze Chow. *Ghost in the (Hollywood) machine: Emergent applications of artificial intelligence in the film industry*. Lug. 2020. DOI: 10.25969/mediarep/14307.
- [14] Troy Yarter. *The Volume: How "The Mandalorian" Revolutionized Filmmaking*. 21 Gen. 2024. URL: <https://illumin.usc.edu/the-volume-how-the-mandalorian-revolutionized-filmmaking/>.
- [15] Bethy Squires. *The Brutalist's AI Controversy, Explained*. 4 Feb. 2025. URL: <https://www.vulture.com/article/the-brutalist-ai-controversy.html>.
- [16] Chris Lee. *How 'The Irishman' Used CGI & Special Effects on Actors*. 9 Gen. 2020. URL: <https://www.vulture.com/2020/01/how-the-irishman-used-cgi-and-special-effects-on-actors.html>.
- [17] *Reply AI Film Festival*. Reply, 2024. URL: <https://www.reply.com/en/artificial-intelligence/reply-ai-film-festival>.

- [18] IULM University. *AI.motion – Festival Italiano del Cinema e dell’Audiovisivo con l’Intelligenza Artificiale*. 2024. URL: <https://www.iulm.it/en/news-ed-eventi/news/ai-motion-festival-italiano-del-cinema-e-audiovisivo-intelligenza-artificiale>.
- [19] Runway AI. *AIFF 2025 – AI Film Festival*. 2025. URL: <https://aiff.runwayml.com/>.
- [20] AI Film Fest. *AI International Film Festival*. 2025. URL: <https://aifilmfest.org/>.
- [21] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. URL: <https://arxiv.org/abs/1301.3781>.
- [22] Sepp Hochreiter e Jürgen Schmidhuber. *Long Short-Term Memory*. 1997. DOI: 10.1162/neco.1997.9.8.1735.
- [23] Ilya Sutskever, Oriol Vinyals e Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. arXiv: 1409.3215. URL: <http://arxiv.org/abs/1409.3215>.
- [24] Ashish Vaswani et al. *Attention Is All You Need*. 2017. URL: <http://arxiv.org/abs/1706.03762>.
- [25] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. URL: <https://arxiv.org/abs/2005.14165>.
- [26] Lech Mazur. *LLM Creative Story-Writing Benchmark*. 2025. URL: <https://github.com/lechmazur/writing>.
- [27] Diederik P Kingma e Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML]. URL: <https://arxiv.org/abs/1312.6114>.
- [28] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: <https://arxiv.org/abs/1406.2661>.
- [29] Kwonjoon Lee et al. *ViTGAN: Training GANs with Vision Transformers*. 2024. arXiv: 2107.04589 [cs.CV]. URL: <https://arxiv.org/abs/2107.04589>.
- [30] Jonathan Ho, Ajay Jain e Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: <https://arxiv.org/abs/2006.11239>.
- [31] Prafulla Dhariwal e Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: 2105.05233 [cs.LG]. URL: <https://arxiv.org/abs/2105.05233>.
- [32] Artificial Analysis. *Text-to-Image Models Leaderboard*. 2025. URL: <https://artificialanalysis.ai/text-to-image>.
- [33] Pengyuan Zhou et al. *A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming*. 2024. arXiv: 2404.16038 [cs.CV]. URL: <https://arxiv.org/abs/2404.16038>.
- [34] Jonathan Ho et al. *Video Diffusion Models*. 2022. arXiv: 2204.03458 [cs.CV]. URL: <https://arxiv.org/abs/2204.03458>.
- [35] William Peebles e Saining Xie. *Scalable Diffusion Models with Transformers*. 2023. arXiv: 2212.09748 [cs.CV]. URL: <https://arxiv.org/abs/2212.09748>.
- [36] Comfy Org. *ComfyUI*. Comfy Org, 2023. URL: <https://www.comfy.org/>.
- [37] Google. *Veo 3*. Google, 2025. URL: <https://veo3ai.org>.
- [38] OpenAI. *Sora*. OpenAI, 2024. URL: <https://sora.chatgpt.com/>.
- [39] Runway AI, Inc. *RunwayML*. Runway AI, Inc., 2023. URL: <https://runwayml.com/>.
- [40] Adobe Inc. *Adobe Firefly*. 2025. URL: <https://firefly.adobe.com/>.
- [41] OpenAI. *MuseNet*. OpenAI, 2019. URL: <https://openai.com/index/musenet/>.
- [42] Adam Roberts et al. *A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music*. 2019. arXiv: 1803.05428 [cs.LG]. URL: <https://arxiv.org/abs/1803.05428>.

- [43] Hao-Wen Dong et al. *MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment*. 2017. arXiv: 1709.06298 [eess.AS]. URL: <https://arxiv.org/abs/1709.06298>.
- [44] Haohe Liu et al. *AudioLDM: Text-to-Audio Generation with Latent Diffusion Models*. 2023. arXiv: 2301.12503 [cs.SD]. URL: <https://arxiv.org/abs/2301.12503>.
- [45] Jesse Engel et al. *DDSP: Differentiable Digital Signal Processing*. 2020. arXiv: 2001.04643 [cs.LG]. URL: <https://arxiv.org/abs/2001.04643>.
- [46] Jonathan Shen et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018. arXiv: 1712.05884 [cs.CL]. URL: <https://arxiv.org/abs/1712.05884>.
- [47] Sander Dieleman Aäron van den Oord. *WaveNet: A Generative Model for Raw Audio*. Accessed: 2025-06-30. DeepMind, set. 2016. URL: <https://deepmind.com/blog/wavenet-generative-model-raw-audio>.
- [48] Chengyi Wang et al. *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers*. 2023. arXiv: 2301.02111 [cs.CL]. URL: <https://arxiv.org/abs/2301.02111>.
- [49] Florian Grötschla et al. “Benchmarking Music Generation Models and Metrics via Human Preference Studies”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr. 2025, pp. 1–5. DOI: 10.1109/icassp49660.2025.10887745. URL: <http://dx.doi.org/10.1109/ICASSP49660.2025.10887745>.
- [50] Suno, Inc. *Suno.AI*. Suno, Inc., 2023. URL: <https://suno.com/>.
- [51] Udio. *Udio*. Udio. 2024. URL: <https://www.udio.com/>.
- [52] Felix Kreuk et al. *AudioGen: Textually Guided Audio Generation*. 2023. arXiv: 2209.15352 [cs.SD]. URL: <https://arxiv.org/abs/2209.15352>.
- [53] ElevenLabs, Inc. *ElevenLabs App Home*. ElevenLabs, Inc., 2023. URL: <https://elevenlabs.io/app/home>.
- [54] Artificial Analysis, Inc. *Text to Speech Models and Providers Leaderboard*. Artificial Analysis, Inc., 2025. URL: <https://artificialanalysis.ai/text-to-speech>.
- [55] Replicate. *Speech-02-HD*. Replicate, 2025. URL: <https://replicate.com/minimax/speech-02-hd>.
- [56] OpenAI. *TTS-1 HD*. OpenAI, 2024. URL: <https://platform.openai.com/docs/models/tts-1-hd>.
- [57] ElevenLabs. *Turbo v2.5*. ElevenLabs, 2024. URL: <https://elevenlabs.io/blog/introducing-turbo-v2-5>.
- [58] Gisele Tong. *To Dear Me*. 2024. URL: <https://www.youtube.com/watch?v=XRdiCzKjWrc>.
- [59] Mentimeter. *Mentimeter*. Mentimeter, 2014. URL: <https://www.mentimeter.com/>.
- [60] DeepSeek AI. *DeepSeek-R1*. 2024. URL: <https://github.com/deepseek-ai/DeepSeek-R1/blob/main/README.md>.
- [61] Anthropic. *Claude 3.7 Sonnet*. 2024. URL: <https://www.anthropic.com/claude/sonnet>.
- [62] OpenAI. *Introduction to GPT-4o*. 2024. URL: https://cookbook.openai.com/examples/gpt4o/introduction_to_gpt4o.
- [63] Masahiro Mori. *The Uncanny Valley*. Translated by Karl F. MacDorman and Norri Kageki, 2012. 1970. URL: <https://spectrum.ieee.org/the-uncanny-valley>.
- [64] Rudolf Arnheim. *Art and Visual Perception: A Psychology of the Creative Eye*. 1954.

- [65] Akira Kurosawa. *Dreams*. 1990. URL: [https://it.wikipedia.org/wiki/Sogni_\(film\)](https://it.wikipedia.org/wiki/Sogni_(film)).
- [66] Norbert Wolf. *Expressionism (Taschen Basic Art)*. 2004.
- [67] Black Forest Labs. *FLUX.1 Documentation*. 2024. URL: <https://huggingface.co/black-forest-labs/FLUX.1-dev>.
- [68] Recraft AI. *Recraft Documentation*. 2024. URL: <https://www.recraft.ai/blog/recraft-introduces-a-revolutionary-ai-model-that-thinks-in-design-language>.
- [69] Google Cloud Platform. *Imagen3 Documentation*. 2024. URL: https://github.com/GoogleCloudPlatform/generative-ai/blob/main/vision/getting-started/imagen3_editing.ipynb.
- [70] Midjourney. *Midjourney Documentation*. 2024. URL: <https://docs.midjourney.com/hc/en-us/categories/32013335627533>.
- [71] Adobe Inc. *Adobe Photoshop*. 2025. URL: <https://www.adobe.com/it/products/photoshop.html>.
- [72] Blackmagic Design. *DaVinci Resolve*. 2025. URL: <https://www.blackmagicdesign.com/it/products/davinciresolve>.

