

# Analisi e Summarization Automatica di Video Tecnici: Un Approccio Multimodale Applicato a Recensioni di Armi da Fuoco Benelli

Balloni Niccolò, Francesco Concetti

June 2025

## Indice

<b>1</b>	<b>Introduzione</b>	<b>3</b>
1.1	Contesto e motivazione . . . . .	3
1.2	Obiettivi del lavoro . . . . .	3
1.3	Struttura della relazione . . . . .	3
<b>2</b>	<b>Background e Stato dell'arte</b>	<b>5</b>
2.1	Approccio tradizionale all'analisi dei video . . . . .	5
2.2	Emergere di strumenti basati su intelligenza artificiale . . . . .	5
2.3	Stato dell'arte open source e principali criticità . . . . .	6
2.4	Motivazioni e opportunità di progetto . . . . .	6
<b>3</b>	<b>Dati</b>	<b>8</b>
3.1	Origine del materiale video . . . . .	8
3.2	Tipologia e struttura dei contenuti . . . . .	8
3.3	Caratteristiche del dataset . . . . .	9
3.4	Annotazione manuale e costruzione del ground truth . . . . .	9
<b>4</b>	<b>Metodi</b>	<b>10</b>
4.1	Estrattore di Clip Salienti: UniVTG . . . . .	10
4.1.1	Architettura del modello . . . . .	10
4.1.2	Pipeline operativa . . . . .	11
4.2	Sintesi testuale con Video-LLaVA . . . . .	12
4.2.1	Architettura e funzionamento del modello . . . . .	12
4.2.2	Rilevanza sulla componente audio . . . . .	12
4.2.3	Strategia di inferenza e generazione su larga scala . . . . .	13
4.2.4	Pipeline alternativa basata su Whisper e BLIP . . . . .	13

<b>5</b>	<b>Risultati</b>	<b>15</b>
5.1	Estrazione delle clip salienti con UniVTG . . . . .	15
5.2	Risultati generati da Video-LLaVA . . . . .	15
5.3	Risultati generati da Whisper e BLIP . . . . .	17
<b>6</b>	<b>Confronto tra strategie e modelli</b>	<b>18</b>
6.1	Confronto: Clip salienti vs. Video completo . . . . .	18
6.2	Video-LLaVA vs. Whisper+BLIP . . . . .	20
6.3	Confronto tra strategie multimodali: Video-LLaVA vs. Whis- per+BLIP . . . . .	20
6.4	Conclusione del confronto . . . . .	21
<b>7</b>	<b>Conclusioni e sviluppi futuri</b>	<b>22</b>

# 1 Introduzione

## 1.1 Contesto e motivazione

Negli ultimi anni, la quantità di contenuti video disponibili online è cresciuta in modo esponenziale, soprattutto grazie a piattaforme come YouTube. Questa enorme mole di dati offre numerose opportunità per l'analisi automatica, ma presenta anche sfide significative in termini di efficienza nella consultazione e nel reperimento delle informazioni rilevanti. In particolare, nei video di tipo tecnico o informativo – come quelli dedicati all'illustrazione delle caratteristiche e dell'utilizzo di fucili – gli utenti hanno spesso la necessità di individuare rapidamente i passaggi chiave, senza dover visionare l'intero filmato. La video summarization, ovvero la capacità di estrarre clip salienti e generare riepiloghi testuali, rappresenta pertanto una tecnologia strategica per facilitare l'accesso a questo tipo di contenuti.

## 1.2 Obiettivi del lavoro

L'obiettivo del presente progetto è lo sviluppo di una pipeline di video summarization focalizzata su video YouTube relativi a fucili, articolata in due fasi principali:

1. **Estrazione automatica delle clip salienti:** impiegare il modello open-source disponibile nel repository UniVTG per identificare e salvare segmenti video di particolare rilevanza semantica o visiva.
2. **Inferenza testuale dei riassunti:** selezionare e integrare un modello di Natural Language Processing (NLP) in grado di trasformare le clip estratte – insieme alle trascrizioni audio o ai metadati – in un sommario testuale sintetico e coerente.

## 1.3 Struttura della relazione

La relazione è organizzata come segue:

- **Capitolo 2 – Background e stato dell'arte:** panoramica delle tecniche di video summarization, dei principali modelli di estrazione clip-based e dei modelli di trascrizione e sintesi testuale.
- **Capitolo 3 - Materiali:** descrizione dei canali YouTube selezionati dal committente e degli strumenti utilizzati per il download e la preparazione dei video.
- **Capitolo 4 – Metodi:** descrizione approfondita dei principali moduli della pipeline. In particolare: l'estrazione delle clip salienti tramite UniVTG, la generazione del riassunto testuale mediante Video-LLaVA e la pipeline alternativa basata su Whisper e BLP
- **Capitolo 5 e 6 – Risultati e osservazioni**

- **Capitolo 7 – Conclusioni e sviluppi futuri:** riflessione finale, limiti attuali del sistema, idee per estensioni future, possibilità di deploy o automazione ulteriore.
- **Bibliografia e appendici:** documentazione di riferimento, articoli scientifici e frammenti di codice rilevanti.

## 2 Background e Stato dell'arte

### 2.1 Approccio tradizionale all'analisi dei video

Fino a pochi anni fa (e in molti contesti ancora oggi), l'analisi dei video — ad esempio recensioni o demo tecniche — veniva svolta in modo interamente manuale da operatori nei reparti marketing, comunicazione o supporto tecnico.

- **Modalità di lavoro:** un team di analisti visiona i filmati, prende appunti sui passaggi salienti e redige report testuali.
- **Limiti:**
  - **Lentezza:** occorrono molte ore/uomo per esaminare video di media lunghezza (5–20min).
  - **Soggettività:** ogni analista può interpretare diversamente la rilevanza di un segmento.
  - **Scalabilità ridotta:** con l'aumento del numero di video, il carico di lavoro cresce in modo non lineare e non sostenibile.

Da questa criticità nasce l'esigenza di automazione, per garantire efficienza, coerenza e capacità di gestire grandi volumi di contenuti.

### 2.2 Emergere di strumenti basati su intelligenza artificiale

Con l'avvento dei primi framework di deep learning e di tecniche di computer vision, sono comparsi strumenti in grado di supportare — e in alcuni casi sostituire — l'analisi manuale:

- **Riconoscimento di scene e oggetti:** modelli CNN come Faster R-CNN (Ren et al., 2015) e Mask R-CNN (He et al., 2017) per il rilevamento preciso di bounding-box e segmentazione o architetture più leggere come YOLOv5 (Ultralytics, 2020) o EfficientDet (Tan et al., 2020) per inference in tempo reale su GPU consumer.
- **Analisi audio:** sistemi ASR (Automatic Speech Recognition) open-source come Mozilla DeepSpeech (Hannun et al., 2014) e wav2vec2.0 (Baevski et al., 2020).
- **Video summarization:** i primi tentativi di sintesi video si basano su algoritmi rank-based o clustering dei keyframe. Modelli come VSUMM (Potapov et al., 2014), che utilizza un approccio basato su K-Means per selezionare i frame più rappresentativi, o come SUM-GAN (Mahasseni et al., 2017), che sfrutta GAN per apprendere quali segmenti mantenere, hanno aperto la strada, ma mostrano limiti evidenti in contesti specifici.

Sul fronte open-source, l'offerta di tool per la video summarization rimane limitata e presenta alcune criticità significative, soprattutto quando applicata a domini di nicchia come quello delle armi da fuoco:

## 2.3 Stato dell'arte open source e principali criticità

Sul fronte open-source, l'offerta di tool per la video summarization rimane limitata e presenta alcune caratteristiche problematiche:

- **Complessità d'uso:** molti dei repository disponibili, come quelli basati su modelli GAN (SUM-GAN) o clustering (VSUMM), richiedono competenze avanzate in machine learning, un processo di setup lungo e la configurazione manuale di numerosi iperparametri.
- **Affidabilità variabile:** gli algoritmi attualmente disponibili tendono a operare su dataset generici (es. SumMe, TVSum) e risultano poco efficaci quando applicati a video con scarsa dinamicità visiva o caratterizzati da contenuti tecnici. Il modello VSUMM, ad esempio, basa la selezione su variazioni nei keyframe, risultando inadatto per video in cui il contesto rilevante è dato dall'audio o da azioni poco evidenti visivamente.
- **Modelli supervisionati deboli:** metodi come SUM-GAN o DR-DSN (Zhou et al., 2018), pur introducendo meccanismi più sofisticati (ad esempio il reinforcement learning), mostrano scarsa generalizzazione fuori dai dataset di benchmark, specialmente in domini con lessico o semantica non convenzionali.
- **Richiesta di risorse hardware elevate:** l'esecuzione di modelli basati su CNN o Transformer (come VideoGPT o TimeSformer) su interi video, accompagnata dall'analisi audio e dalla trascrizione, richiede GPU potenti e notevole disponibilità di memoria, rendendo l'utilizzo poco accessibile fuori da contesti accademici o enterprise.

## 2.4 Motivazioni e opportunità di progetto

Alla luce delle limitazioni evidenziate, il nostro progetto si propone di:

1. **Semplificare l'adozione** di tecniche di video summarization open-source, fornendo una pipeline pronta all'uso, facilmente configurabile anche da utenti non esperti.
2. **Ottenere estrazioni più affidabili** nel dominio specifico dei video su fucili, attraverso l'integrazione di segnali multimodali (video, audio e trascrizione) e l'ottimizzazione dei criteri di selezione dei segmenti.
3. **Ridurre i requisiti hardware**, sfruttando tecniche di inferenza efficienti, sistemi di frame selection adattiva e meccanismi di pre-elaborazione che minimizzano il carico computazionale.

Un ulteriore aspetto critico, particolarmente rilevante nel nostro dominio applicativo, è la notevole variabilità dei contesti e dei momenti salienti all'interno di ciascun video. Nei contenuti analizzati, si alternano ad esempio scene statiche

in cui l'utente è seduto alla scrivania e spiega verbalmente, a sequenze dinamiche all'aperto con dimostrazioni pratiche. Questa eterogeneità rende inefficace l'utilizzo di approcci unimodali e giustifica la necessità di strategie multimodali complementari, capaci di integrare informazioni visive, uditive e testuali per una selezione più robusta e adattiva dei segmenti rilevanti.

In questo modo si offre uno strumento riproducibile, scalabile e concretamente utilizzabile da tutor universitari, ricercatori e sviluppatori, in grado di facilitare l'analisi rapida di video tecnici. Nel prossimo capitolo verranno descritte le caratteristiche del dataset selezionato, le fonti YouTube e le operazioni di pre-processing preliminari alla pipeline di estrazione delle clip.

## 3 Dati

Per l’addestramento e la valutazione della pipeline progettata, è stato costruito un dataset su misura composto da contenuti video specifici del dominio armiero, in particolare recensioni tecniche di armi da fuoco. La raccolta e selezione del materiale è avvenuta in stretta collaborazione con l’azienda partner Benelli Armi.

### 3.1 Origine del materiale video

Benelli Armi ha fornito una lista di 26 canali YouTube, accuratamente selezionati per l’elevata qualità e pertinenza dei contenuti pubblicati nel settore delle armi sportive e da caccia. Da questi canali, il team ha individuato e scaricato 10 video rappresentativi, scelti manualmente in base ai seguenti criteri:

- chiarezza del parlato;
- presenza di contenuti tecnici espliciti;
- copertura di diversi formati e ambienti di ripresa (indoor/outdoor);
- lunghezza compatibile con i limiti computazionali della pipeline.

### 3.2 Tipologia e struttura dei contenuti

Tutti i video appartengono al genere delle recensioni tecniche, un formato ampiamente diffuso nella comunicazione commerciale e informativa del settore armiero. Questi video si caratterizzano per una struttura ricorrente e ben definita, articolata in tre segmenti principali:

- **Presentazione dell’arma da parte di un esperto:** l’autore introduce il modello oggetto della recensione, illustrandone le principali caratteristiche costruttive, come il tipo di meccanismo (es. inerziale o a gas), i materiali, la finitura superficiale, la configurazione del calcio e il sistema di mira. Spesso vengono inquadrati i dettagli dell’arma appoggiata su un banco da lavoro o esibita manualmente.
- **Test pratici sul campo:** il recensore passa alla dimostrazione funzionale dell’arma, effettuando prove di tiro in ambienti esterni, solitamente in poligoni di tiro o campi da caccia. In queste sezioni il contenuto visivo è più dinamico e il parlato si accompagna spesso a commenti estemporanei, considerazioni soggettive e descrizioni “live” dell’esperienza d’uso.
- **Confronti o conclusioni tecniche:** alcuni video presentano confronti con altri modelli simili, o riepiloghi delle osservazioni raccolte, offrendo una valutazione complessiva dell’arma sotto diversi aspetti (ergonomia, precisione, gestione del rinculo, maneggevolezza, ecc.).



Questa eterogeneità di contesti comunicativi — alternando momenti statici in ambienti controllati a fasi dinamiche sul campo — rappresenta una sfida significativa per i modelli di video summarization, giustificando l'impiego di approcci multimodali in grado di integrare visione, linguaggio e suono.

### 3.3 Caratteristiche del dataset

Il dataset finale è composto da 10 video con le seguenti durate (in minuti): 7, 26, 19, 11, 13, 10, 20, 9, 11 e 10, per una durata media complessiva di circa 13,6 minuti.

I video sono stati utilizzati nel formato originale senza interventi di pre-processing. La pipeline di analisi si occupa direttamente della loro elaborazione, partendo dalla versione integra e automatizzando l'estrazione dei segmenti rilevanti tramite il modello UniVTG (cfr. Capitolo 4).

### 3.4 Annotazione manuale e costruzione del ground truth

Per poter valutare in modo accurato le prestazioni dei modelli di estrazione e generazione testuale, è stato realizzato un ground truth manuale. Ciascun membro del team ha contribuito ad annotare i video selezionando, per ogni contenuto, i segmenti ritenuti "informativamente" salienti.

A ciascuna clip così individuata è stato associato un insieme di domande guida, progettate per stimolare la generazione automatica di contenuti e fornire una base coerente per il confronto tra i diversi modelli di sintesi.

```
# Domande specifiche per video di recensione di fucili
questions = [
    "Cosa succede in questa clip?",
    "Che modello di fucile viene mostrato o recensito?",
    "Viene descritto qualche dettaglio tecnico del fucile? Se sì, quali?",
    "Viene mostrato il funzionamento pratico del fucile? Ad esempio, il caricamento, lo sparo o il meccanismo interno?",
    "Ci sono commenti o opinioni sulle prestazioni del fucile? Se sì, quali?",
    "Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?",
    "Vengono mostrati test di tiro o prove pratiche? Se sì, su quali bersagli e con quali risultati?",
    "Viene fatto un confronto con altri modelli di fucile o armi simili?",
    "Ci sono indicazioni sull'utilizzo previsto del fucile? (Es. caccia, tiro sportivo, softair, difesa, collezionismo)",
    "L'utente parla di pregi e difetti? Se sì, quali vengono evidenziati?",
]
```

Figure 1: Schermata delle domande

## 4 Metodi

La pipeline proposta per la video summarization si articola in due componenti principali: l'estrazione automatica delle clip salienti da video completi e la generazione di descrizioni testuali a partire da tali clip. Le due fasi lavorano in sequenza e sono concepite per operare in modo modulare, facilitando la sostituzione o l'adattamento dei modelli in base al contesto operativo o alle risorse computazionali disponibili.

Per l'estrazione delle clip è stato adottato UniVTG, un modello open-source sviluppato per il video-language temporal grounding e adattabile al task di highlight detection. Per la sintesi testuale, sono stati valutati due approcci: un modello multimodale avanzato (Video-LLaVA) e una soluzione alternativa più leggera e interpretabile, basata su Whisper per la trascrizione vocale e BLIP per la generazione di caption visive.

### 4.1 Estrattore di Clip Salienti: UniVTG

La prima fase della pipeline progettata riguarda l'identificazione automatica delle porzioni più informative all'interno di ciascun video, con l'obiettivo di ridurre la lunghezza dei contenuti da elaborare mantenendo un alto livello di rilevanza semantica. A questo scopo è stato utilizzato UniVTG (Unified Video-Text Grounding), un modello open-source progettato per associare input video a descrizioni testuali, e capace di svolgere diversi task: moment retrieval, highlight detection e video summarization.

#### 4.1.1 Architettura del modello

Il sistema si basa su un'architettura a componenti specializzati, progettati per apprendere la corrispondenza tra contenuti video e descrizioni testuali:

- **Encoder video:** implementato tramite una rete 3D convoluzionale (R3D-18) o, in alternativa, un Swin Transformer 3D, questo modulo riceve in input brevi clip video suddivise in sottosequenze (tipicamente 16 frame) e produce per ciascuna un embedding vettoriale che rappresenta la dinamica spazio-temporale.
- **Encoder testuale:** basato sul CLIP Text Encoder (ViT-B/32), consente di generare una rappresentazione semantica del contenuto testuale, come un prompt descrittivo o una trascrizione associata al video.
- **Allineamento multimodale:** i due flussi (video e testo) vengono proiettati in uno spazio latente comune, mediante layer fully-connected. L'interazione avviene tramite concatenazione o cross-attention, consentendo al modello di valutare la corrispondenza tra contenuti visivi e concetti linguistici.
- **Transformer temporale:** un modulo Transformer elabora la sequenza di embedding combinati, modellando le dipendenze temporali tra i segmenti e includendo meccanismi di positional encoding.

- **Heads di output:**

- Interval Head: predice coppie di timestamp (inizio-fine) corrispondenti a possibili intervalli informativi.
- Saliency Head: assegna a ciascun segmento un valore continuo che esprime il grado di salienza.
- Highlight Curve Head (opzionale): restituisce una curva di attenzione distribuita sulla timeline.

#### 4.1.2 Pipeline operativa

L'inferenza avviene seguendo una sequenza di passaggi standardizzati:

1. **Setup:** clonazione del repository e installazione dipendenze (timm, transformers, decord, ffmpeg-python, ecc.) e caricamento del checkpoint pre-allenato (.ckpt) in GPU/CPU.
2. **Analisi video preliminare:** viene calcolata la durata totale del video, sulla base della quale si imposta una soglia massima di contenuto da processare (nel nostro caso il 25% della durata). Questa scelta è stata effettuata per limitare il carico computazionale, concentrando l'analisi sulle parti iniziali e centrali dei video, dove si concentra generalmente la densità informativa più elevata.
3. **Campionamento e pre-elaborazione visiva:** i frame vengono estratti a un framerate ridotto di 0.5 fps, valore deciso per ottenere una copertura visiva rappresentativa ma sostenibile in termini computazionali. Ogni frame viene ridimensionato a  $224 \times 224$  pixel e raggruppato in segmenti.
4. **Estrazione delle feature video:** le clip vengono processate tramite l'encoder visivo (R3D-18), che restituisce una sequenza di vettori spaziali e temporali, rappresentativi del contenuto dinamico.
5. **Iniezione del prompt testuale:** un prompt generico ("Generate a concise summary of the following video") viene tokenizzato e trasformato in embedding testuali tramite il CLIP encoder.
6. **Inferenza multimodale:** le due sequenze vengono fuse nel Transformer, che restituisce un vettore di saliency scores: un punteggio di rilevanza per ciascun segmento.
7. **Selezione delle clip:** i segmenti con i punteggi di salienza più elevati vengono ordinati in modo decrescente. Per evitare sovrapposizioni temporali tra clip adiacenti, viene applicato un filtro basato sulla distanza minima tra segmenti, espressa in frame. In particolare, viene scartato ogni frame che si trovi a una distanza inferiore a un valore soglia (min\_frame\_distance) da un frame già selezionato.

8. **Estrazione delle clip:** le clip vengono infine estratte centrando ciascun segmento attorno al frame selezionato, utilizzando una finestra temporale della durata specificata (es. 8–15 secondi). Il risultato è un insieme di clip brevi, semanticamente dense, pronte per la fase successiva di analisi testuale.

## 4.2 Sintesi testuale con Video-LLaVA

Una volta completata l'estrazione delle clip salienti tramite UniVTG, la seconda fase della pipeline prevede la generazione automatica di una descrizione testuale per ciascun segmento. Per questo scopo è stato adottato Video-LLaVA, un sistema multimodale avanzato progettato per combinare input visivi e linguistici e generare risposte coerenti in linguaggio naturale.

### 4.2.1 Architettura e funzionamento del modello

Video-LLaVA estende un Large Language Model (LLM), in particolare Vicuna-7B, con capacità visive, consentendogli di comprendere e descrivere contenuti video. L'architettura si basa su un approccio chiamato *Alignment Before Projection*<sup>1</sup>, derivato dal framework LanguageBind, e si articola in tre componenti principali:

1. **Encoder visivo pre-allineato:** basato su OpenCLIP, è stato addestrato su coppie video-testo e genera una sequenza di embedding visivi per ciascun frame selezionato.
2. **Modulo Q-former multimodale:** applica meccanismi di cross-attention tra un set di query apprendibili e le feature visive, producendo un numero limitato di token visuali semanticamente ricchi.
3. **Decoder autoregressivo (LLM):** riceve in input sia le trascrizioni testuali (o prompt) sia i token visivi e, tramite attenzione incrociata, genera una risposta descrittiva o un riassunto della clip. Questo approccio consente al modello di integrare informazioni linguistiche e visive in modo coeso.

### 4.2.2 Rilevanza sulla componente audio

Un aspetto particolarmente importante nel nostro caso d'uso è la presenza di contenuti sonori informativi, come parlato, rumori ambientali e suoni di interazione con gli oggetti. I modelli come Video-LLaVA sono in grado di elaborare anche la componente audio dei video, sia come input diretto (quando supportato) sia come informazione implicitamente integrata nei pretraining multimodali.

---

<sup>1</sup>Il principio di Alignment Before Projection consiste nell'allineare prima le modalità visive e linguistiche in uno spazio latente comune, mantenendo intatte le caratteristiche di ciascuna, e solo in un secondo momento proiettare tali rappresentazioni in un formato compatibile con il modello linguistico. Questo approccio permette di preservare la ricchezza semantica dei token visivi, migliorando la coerenza tra l'input video e l'output testuale.

In particolare, nel dominio trattato (recensioni e test di armi), i segnali acustici forniscono un contributo fondamentale alla comprensione della scena: ad esempio, la presenza di un suono di sparo, di un rimbalzo metallico o della voce dell'operatore permette al sistema di inferire il contesto d'azione anche in assenza di cambiamenti visivi marcati.

Questa capacità risulta essenziale per interpretare correttamente i contenuti tecnici e rende i modelli multimodali un'ottima scelta per il tipo di video analizzati nel progetto.

### 4.2.3 Strategia di inferenza e generazione su larga scala

L'intero processo di generazione testuale è stato progettato per operare in modalità batch su più clip in parallelo, così da garantire efficienza e ripetibilità nella fase di analisi. A ciascuna clip viene associato automaticamente l'insieme di domande definite in fase di annotazione (si veda la sezione precedente), con l'obiettivo di guidare il modello nella produzione di descrizioni dettagliate, coerenti e strutturate.

Le risposte generate per ogni clip vengono salvate e successivamente sintetizzate in una descrizione complessiva dell'intero contenuto video, tramite un modello linguistico esterno. Questo approccio consente di gestire in modo scalabile l'analisi di grandi volumi di dati, mantenendo l'intervento umano limitato alla sola fase di revisione.<sup>2</sup>

### 4.2.4 Pipeline alternativa basata su Whisper e BLIP

Durante la fase di sviluppo, è stata parallelamente progettata una soluzione alternativa e più leggera per l'analisi testuale delle clip video, con l'obiettivo di garantire la flessibilità della pipeline anche in contesti con risorse computazionali più limitate. Tale approccio si basa sull'utilizzo combinato di Whisper (per la trascrizione audio) e BLIP (per la descrizione visiva), ed è stato implementato e testato su Google Colab.

Questa variante è stata pensata per semplificare la struttura operativa e offrire una modalità modulare di estrazione del contenuto semantico, mantenendo comunque un buon livello di interpretabilità e copertura informativa.

Il funzionamento della pipeline si articola in più fasi complementari. Per ogni video, viene effettuata un'estrazione regolare di un numero limitato di fotogrammi (tipicamente sei), distribuiti uniformemente lungo l'intera durata. Questo consente di ottenere una copertura visiva rappresentativa senza dover elaborare ogni frame, mantenendo basso il carico computazionale.

Ciascun fotogramma viene analizzato tramite BLIP (Bootstrapped Language Image Pretraining), un modello pre-addestrato che genera descrizioni

---

<sup>2</sup>L'esecuzione della pipeline è avvenuta su un'infrastruttura locale dotata di GPU ad alte prestazioni, scelta necessaria per garantire la stabilità e la compatibilità durante l'inferenza dei modelli di grandi dimensioni (come Video-LLaVA-7B). Le problematiche affrontate durante la fase sperimentale verranno descritte nel dettaglio nei capitoli successivi.

testuali sintetiche e semanticamente rilevanti del contenuto visivo. Le caption ottenute vengono poi filtrate per rimuovere ridondanze o descrizioni troppo simili, ottenendo così un insieme conciso ma informativamente completo.

Parallelamente, la traccia audio viene estratta e trascritta tramite Whisper, un modello di tipo speech-to-text che permette di convertire in modo affidabile il parlato in testo. Questo passaggio è particolarmente rilevante per i video analizzati, nei quali il commento vocale, le istruzioni tecniche e i rumori ambientali (es. spari, interazioni con oggetti metallici) costituiscono una fonte primaria di informazione semantica.

Il risultato finale consiste in un output testuale per ciascuna clip, composto da due elementi distinti:

- una serie di descrizioni visive sintetiche prodotte da BLIP;
- una trascrizione completa della componente audio ottenuta con Whisper.

Sebbene questa pipeline non integri simultaneamente le due modalità come avviene in Video-LLaVA, rappresenta comunque una soluzione solida, interpretabile e facilmente adattabile, particolarmente indicata per scenari a risorse ridotte. In quest’ottica, costituisce un modulo complementare all’interno dell’intera architettura progettuale, offrendo una via alternativa per garantire l’estrazione di informazioni testuali anche in assenza di modelli multimodali complessi.

## 5 Risultati

Questa sezione illustra concretamente gli output generati dalle pipeline sperimentali descritte nei capitoli precedenti. L'obiettivo è valutare la qualità e l'utilità delle informazioni estratte dalle clip video tramite modelli diversi, mettendo a confronto soluzioni multimodali avanzate e approcci più leggeri ma interpretabili.

Partiremo dall'estrazione automatica delle clip salienti con UniVTG, per poi analizzare i risultati testuali prodotti da Video-LLaVA e, infine, da Whisper + BLIP. Verranno inoltre presentati alcuni confronti qualitativi tra i modelli, e motivate le scelte metodologiche adottate, alla luce dell'efficacia delle pipeline testate.

### 5.1 Estrazione delle clip salienti con UniVTG

UniVTG è stato utilizzato per identificare automaticamente i momenti a maggiore rilevanza semantica all'interno dei video. Il modello calcola uno *score di salienza* per ciascun segmento video e restituisce un insieme di clip corrispondenti ai picchi massimi nella curva di attenzione multimodale.

Nel caso del **Video 6**, ad esempio, sono stati selezionati i **Top-7 clip indices**, ovvero i 7 intervalli con maggiore probabilità di contenere informazioni salienti (Figura 2). Questi indici rappresentano i punti centrali delle clip generate, e vengono successivamente utilizzati per l'analisi testuale.

```
pytorch_model bin: 100% 599M/599M [00:07<00:00, 142MB/s]
Top-7 clip indices (filtrati): [np.int64(88), np.int64(110), np.int64(299), np.int64(235), np.int64(67), np.int64(41), np.int64(252)]
model callensors: 100% 599M/599M [00:06<00:00, 121MB/s]
Clip estratti e salvati in: /content/drive/MyDrive/clip_output_univtg/video6
```

Figure 2: Output generato da UniVTG: indici dei 7 segmenti più rilevanti del Video 6, selezionati secondo la curva di salienza.

A partire da questi timestamp, ogni video viene suddiviso in clip autonome e semanticamente informative. La Figura 3 mostra un esempio visivo di alcune clip selezionate.



Figure 3: Frame centrali delle clip estratte da UniVTG per il Video 6. Ogni immagine rappresenta un punto ad alta salienza multimodale.

### 5.2 Risultati generati da Video-LLaVA

Il sistema Video-LLaVA restituisce per ciascuna clip video un insieme di risposte a dieci domande predefinite, progettate per ottenere una comprensione strut-

turata e dettagliata del contenuto multimodale (visivo e audio). Le risposte vengono salvate in un file `.json`, in cui ogni domanda è mappata alla relativa risposta per ogni clip processata. Un esempio concreto dell'output è mostrato in Figura 4.

```
"clip_00.mp4": {
  "Cosa succede in questa clip?": "In questa clip, un uomo spara un colpo con una pistola e poi spara altre due palle.",
  "Che modello di fucile viene mostrato o recensito?": "Il fucile viene mostrato e recensito in dettaglio.",
  "Viene descritto qualche dettaglio tecnico del fucile? Se sì, quali?": "Sì, il fucile viene descritto in dettaglio e viene menzionato che è un fucile di precisione.",
  "Viene mostrato il funzionamento pratico del fucile? Ad esempio, il caricamento, lo sparo o il meccanismo interno?": "Sì, il fucile viene mostrato in azione e viene descritto il funzionamento pratico.",
  "Ci sono commenti o opinioni sulle prestazioni del fucile? Se sì, quali?": "Sì, il fucile viene descritto come molto preciso e potente.",
  "Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?": "Sì, il fucile viene mostrato con un accessorio che non viene specificato.",
  "Vengono mostrati test di tiro o prove pratiche? Se sì, su quali bersagli e con quali risultati?": "Sì, il fucile viene mostrato in azione su un bersaglio e viene descritto come molto preciso.",
  "Viene fatto un confronto con altri modelli di fucile o armi simili?": "Sì, il fucile viene paragonato ad altri modelli di fucile di precisione.",
  "Ci sono indicazioni sull'utilizzo previsto del fucile? (Es. caccia, tiro sportivo, softair, difesa, collezionismo)": "Sì, il fucile viene descritto come ideale per il tiro sportivo e la caccia.",
  "L'utente parla di pregi e difetti? Se sì, quali vengono evidenziati?": "Sì, l'utente parla dei pregi e dei difetti del fucile e li evidenzia in modo dettagliato.",
},
"clip_01.mp4": {
  "Cosa succede in questa clip?": "In questa clip, un uomo in un campo di tiro spara una pistola AR-47 e parla all'interno del campo.",
  "Che modello di fucile viene mostrato o recensito?": "Il fucile AR-47 viene mostrato e recensito in questa clip.",
  "Viene descritto qualche dettaglio tecnico del fucile? Se sì, quali?": "Sì, il fucile viene descritto come un'arma potente e precisa. Viene menzionato che il fucile ha un'ottima precisione e che è molto popolare.",
  "Viene mostrato il funzionamento pratico del fucile? Ad esempio, il caricamento, lo sparo o il meccanismo interno?": "No, non viene mostrato il funzionamento pratico del fucile. La clip si concentra principalmente.",
  "Ci sono commenti o opinioni sulle prestazioni del fucile? Se sì, quali?": "Sì, il fucile viene descritto come un'arma potente e precisa. Viene menzionato che il fucile ha un'ottima precisione e che è molto popolare.",
  "Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?": "No, non sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile. La clip si concentra principalmente.",
  "Vengono mostrati test di tiro o prove pratiche? Se sì, su quali bersagli e con quali risultati?": "No, non vengono mostrati test di tiro o prove pratiche. La clip si concentra principalmente sulla presentazione.",
  "Viene fatto un confronto con altri modelli di fucile o armi simili?": "No, non viene fatto alcun confronto con altri modelli di fucile o armi simili. La clip si concentra principalmente sulla presentazione e.",
  "Ci sono indicazioni sull'utilizzo previsto del fucile? (Es. caccia, tiro sportivo, softair, difesa, collezionismo)": "No, non ci sono indicazioni sull'utilizzo previsto del fucile. La clip si concentra principalmente.",
  "L'utente parla di pregi e difetti? Se sì, quali vengono evidenziati?": "No, non viene menzionato alcun pregio o difetto del fucile. La clip si concentra principalmente sulla presentazione e la descrizione del.",
},
"clip_02.mp4": {
  "Cosa succede in questa clip?": "In questa clip, un uomo in un campo di tiro spara una pistola e poi spara un'altra pistola.",
  "Che modello di fucile viene mostrato o recensito?": "Il fucile viene mostrato e recensito in dettaglio.",
  "Viene descritto qualche dettaglio tecnico del fucile? Se sì, quali?": "Sì, il fucile viene descritto in dettaglio e viene mostrata in azione.",
  "Viene mostrato il funzionamento pratico del fucile? Ad esempio, il caricamento, lo sparo o il meccanismo interno?": "Sì, il funzionamento pratico del fucile viene mostrato in azione.",
  "Ci sono commenti o opinioni sulle prestazioni del fucile? Se sì, quali?": "Sì, ci sono commenti e opinioni sulle prestazioni del fucile.",
  "Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?": "Sì, vengono mostrati accessori, modifiche e personalizzazioni del fucile.",
  "Vengono mostrati test di tiro o prove pratiche? Se sì, su quali bersagli e con quali risultati?": "Sì, vengono mostrati test di tiro o prove pratiche.",
  "Viene fatto un confronto con altri modelli di fucile o armi simili?": "Sì, il fucile viene comparato con altri modelli di fucile o armi simili.",
  "Ci sono indicazioni sull'utilizzo previsto del fucile? (Es. caccia, tiro sportivo, softair, difesa, collezionismo)": "Sì, ci sono indicazioni sull'utilizzo previsto del fucile.",
  "L'utente parla di pregi e difetti? Se sì, quali vengono evidenziati?": "Sì, l'utente parla di pregi e difetti del fucile.",
},
"clip_03.mp4": {
  "Cosa succede in questa clip?": "In questa clip, un uomo con un'arma da fuoco in mano parla e spara un proiettile in una bara di legno.",
  "Che modello di fucile viene mostrato o recensito?": "Il video non fornisce informazioni specifiche sul modello di fucile mostrato o recensito.",
  "Viene descritto qualche dettaglio tecnico del fucile? Se sì, quali?": "Il video non fornisce informazioni specifiche sui dettagli tecnici del fucile.",
  "Viene mostrato il funzionamento pratico del fucile? Ad esempio, il caricamento, lo sparo o il meccanismo interno?": "Il video non fornisce informazioni specifiche sul funzionamento pratico del fucile.",
  "Ci sono commenti o opinioni sulle prestazioni del fucile? Se sì, quali?": "Il video non fornisce commenti o opinioni sulle prestazioni del fucile.",
  "Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?": "Il video non fornisce informazioni specifiche sui accessori, modifiche o personalizzazioni del fucile.",
  "Vengono mostrati test di tiro o prove pratiche? Se sì, su quali bersagli e con quali risultati?": "Il video non fornisce informazioni specifiche sui test di tiro o le prove pratiche effettuate con il fucile.",
  "Viene fatto un confronto con altri modelli di fucile o armi simili?": "Il video non fornisce informazioni specifiche sul confronto con altri modelli di fucile o armi simili.",
  "Ci sono indicazioni sull'utilizzo previsto del fucile? (Es. caccia, tiro sportivo, softair, difesa, collezionismo)": "Il video non fornisce informazioni specifiche sull'utilizzo previsto del fucile.",
  "L'utente parla di pregi e difetti? Se sì, quali vengono evidenziati?": "Il video non fornisce informazioni specifiche sui pregi e difetti evidenziali dall'utente.",
}
```

Figure 4: Estratto del file `.json` prodotto da Video-LLaVA, contenente le risposte alle dieci domande per ciascuna clip.

In parallelo, il sistema genera anche un file `.txt` contenente un prompt che richiede la sintesi delle informazioni ottenute da tutte le clip analizzate. Il prompt è pensato per essere processato da un LLM esterno (es. ChatGPT), con l'obiettivo di fornire un riassunto unico e coerente dell'intero contenuto video.

```

Il fornirò una serie di domande, seguite dalle risposte ottenute dall'analisi automatica di diverse clip estratte dallo stesso video.
Il tuo compito è leggere tutte le risposte fornite e rispondere nuovamente a queste stesse domande, sintetizzando e combinando le informazioni provenienti da tutte le clip, in modo da fornire delle risposte più complete, accurate e
affidabili rispetto a quanto osservato nell'intero video.

...

• Clip: clip_00.mp4
Domanda: Cosa succede in questa clip?
Risposta: In questa clip, un uomo spara un colpo con una pistola e poi spara altre due palle.
Domanda: Che modello di fucile viene mostrato o recensito?
Risposta: Il fucile viene mostrato e recensito in dettaglio.
Domanda: Viene descritto qualche dettaglio tecnico del fucile? Se sì, quali?
Risposta: Sì, il fucile viene descritto in dettaglio e viene menzionato che è un fucile di precisione.
Domanda: Viene mostrato il funzionamento pratico del fucile? Ad esempio, il caricamento, lo sparo o il meccanismo interno?
Risposta: Sì, il fucile viene mostrato in azione e viene descritto il funzionamento pratico.
Domanda: Ci sono commenti o opinioni sulle prestazioni del fucile? Se sì, quali?
Risposta: Sì, il fucile viene descritto come molto preciso e potente.
Domanda: Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?
Risposta: Sì, il fucile viene mostrato con un accessorio che non viene specificato.
Domanda: Vengono mostrati test di tiro o prove pratiche? Se sì, su quali bersagli e con quali risultati?
Risposta: Sì, il fucile viene mostrato in azione su un bersaglio e viene descritto come molto preciso.
Domanda: Viene fatto un confronto con altri modelli di fucile o armi simili?
Risposta: Sì, il fucile viene paragonato ad altri modelli di fucile di precisione.
Domanda: Ci sono indicazioni sull'utilizzo previsto del fucile? (Es. caccia, tiro sportivo, softair, difesa, collezionismo)
Risposta: Sì, il fucile viene descritto come ideale per il tiro sportivo e la caccia.
Domanda: L'utente parla di pregi e difetti? Se sì, quali vengono evidenziati?
Risposta: Sì, l'utente parla dei pregi e dei difetti del fucile e li evidenzia in modo dettagliato.
...

```

Figure 5: Parte iniziale del prompt generato automaticamente nel file `.txt`, che guida un LLM alla sintesi dei contenuti.



```

...
🔴 Ora rispondi nuovamente alle seguenti domande, tenendo conto di tutte le informazioni fornite sopra, come se avessi visto il video completo.
Per favore fornisci la tua risposta seguendo questa struttura:

Domanda: Cosa succede in questa clip?
Risposta:

Domanda: Che modello di fucile viene mostrato o recensito?
Risposta:

Domanda: Viene descritto qualche dettaglio tecnico del fucile? Se sì, quali?
Risposta:

Domanda: Viene mostrato il funzionamento pratico del fucile? Ad esempio, il caricamento, lo sparo o il meccanismo interno?
Risposta:

Domanda: Ci sono commenti o opinioni sulle prestazioni del fucile? Se sì, quali?
Risposta:

Domanda: Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?
Risposta:

Domanda: Vengono mostrati test di tiro o prove pratiche? Se sì, su quali bersagli e con quali risultati?
Risposta:

Domanda: Viene fatto un confronto con altri modelli di fucile o armi simili?
Risposta:

Domanda: Ci sono indicazioni sull'utilizzo previsto del fucile? (es. caccia, tiro sportivo, softair, difesa, collezionismo)
Risposta:

Domanda: L'utente parla di pregi e difetti? Se sì, quali vengono evidenziati?
Risposta:

```

Figure 6: Seconda parte del prompt, con la lista delle domande a cui il modello deve rispondere sintetizzando i dati.

Il risultato finale è un testo riassuntivo generato da un modello linguistico a partire dalle informazioni raccolte clip per clip. Questo testo rappresenta una visione unificata, interpretativa e compatta del contenuto video originale. Un esempio è mostrato in Figura 7.

```

Domanda: Cosa succede in questa clip?
Risposta: Nel video complessivo, si alternano diverse scene in cui un uomo mostra, recensisce e testa vari tipi di armi da fuoco, tra cui fucili di precisione, un AK-47 e un'arma a canna mozza. Vengono effettuati spari su diversi bersagli, tra cui bersagli classici, bare di legno o fango, e si osservano sia descrizioni teoriche che dimostrazioni pratiche delle armi.

Domanda: Che modello di fucile viene mostrato o recensito?
Risposta: Vengono mostrati e recensiti diversi modelli di fucile, tra cui un fucile di precisione, un AK-47 e un fucile a canna mozza. Alcuni fucili sono dotati di accessori come silenziatori e sistemi di mira personalizzabili.

Domanda: Viene descritto qualche dettaglio tecnico del fucile? Se sì, quali?
Risposta: Sì, vengono descritti dettagli tecnici come la precisione e la potenza del fucile di precisione, il sistema di munizione singola, la presenza di un sistema di mira (mirabile), la struttura con impugnatura in legno e asta in metallo nel fucile a canna mozza, e le caratteristiche generali dell'AK-47 in termini di popolarità e precisione.

Domanda: Viene mostrato il funzionamento pratico del fucile? Ad esempio, il caricamento, lo sparo o il meccanismo interno?
Risposta: In alcune clip viene mostrato il funzionamento pratico, inclusi il caricamento, lo sparo e in certi casi il meccanismo interno. In altre, come quella dell'AK-47 o del fucile a canna mozza, la dimostrazione pratica è limitata o assente.

Domanda: Ci sono commenti o opinioni sulle prestazioni del fucile? Se sì, quali?
Risposta: Sì, l'utente esprime opinioni positive riguardo la precisione e la potenza dei fucili, sottolineando che sono adatti per il tiro sportivo, la caccia e la difesa personale. Viene inoltre menzionata la popolarità dell'AK-47 tra i collezionisti.

Domanda: Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?
Risposta: Sì, alcuni fucili sono mostrati con accessori come silenziatori, sistemi di mira regolabili e altre modifiche personalizzabili, che ne migliorano le prestazioni e l'adattabilità alle esigenze dell'utilizzatore.

Domanda: Vengono mostrati test di tiro o prove pratiche? Se sì, su quali bersagli e con quali risultati?
Risposta: Sì, vengono mostrati test di tiro su vari tipi di bersagli, inclusi bersagli statici, bersagli in movimento e bare di legno o fango. I risultati mostrano una buona precisione e potenza delle armi testate.

Domanda: Viene fatto un confronto con altri modelli di fucile o armi simili?
Risposta: Sì, in alcune clip vengono fatti confronti tra diversi modelli di fucile, valutandone prestazioni, design e prezzo.

Domanda: Ci sono indicazioni sull'utilizzo previsto del fucile? (es. caccia, tiro sportivo, softair, difesa, collezionismo)
Risposta: Sì, viene indicato che i fucili sono adatti principalmente per il tiro sportivo, la caccia e la difesa personale, mentre l'AK-47 è menzionato anche come arma apprezzata dai collezionisti.

Domanda: L'utente parla di pregi e difetti? Se sì, quali vengono evidenziati?
Risposta: Sì, vengono discussi pregi come l'elevata precisione, potenza e qualità costruttiva. Tra i difetti, si evidenziano il prezzo elevato di alcuni modelli e la complessità del caricamento di alcune armi. Alcune clip non approfondiscono invece questa analisi.

```

Figure 7: Output finale generato da un LLM esterno a partire dal prompt costruito con le risposte raccolte da Video-LLaVA.

### 5.3 Risultati generati da Whisper e BLIP

Come alternativa leggera a Video-LLaVA, la combinazione dei modelli Whisper e BLIP consente di ottenere informazioni testuali da ciascuna clip tramite una pipeline più modulare, adatta anche ad ambienti con risorse computazionali limitate.

Il processo produce, per ogni clip, una descrizione testuale basata su due fonti:

- La trascrizione automatica dell'audio effettuata da **Whisper**, che permette di recuperare i contenuti verbali (es. spiegazioni, commenti, riferimenti tecnici).
- Le descrizioni generate da **BLIP** a partire da fotogrammi selezionati, che forniscono informazioni visive sintetiche.

Ogni clip è stata analizzata rispondendo a un set fisso di dieci domande, le stesse utilizzate per Video-LLaVA, permettendo così un confronto diretto tra le due pipeline. Le risposte sono raccolte in un file di output strutturato, come mostrato in Figura 8.

```
Clip: clip_00.mp4
Domanda 1: Un uomo in camicia nera e cappello tiene e mostra un fucile, parla di un caricatore a tamburo per 1903 con capacità di circa 30 colpi.
Domanda 2: Viene menzionato il modello M2.
Domanda 3: Sì, viene descritto un caricatore a tamburo (dopo negative) che si aggancia alla M2, probabilmente con caricamento automatico o elicoidale. Capacità circa 30 colpi.
Domanda 4: No, non si descrive il funzionamento pratico dettagliato, solo il caricatore.
Domanda 5: Implicito un commento positivo sul caricatore e sull'operabilità di sparare con esso, senza problemi durante l'uso.
Domanda 6: Sì, è mostrato un caricatore a tamburo come accessorio.
Domanda 7: Non sono menzionati test di tiro o risultati specifici.
Domanda 8: No, nessun confronto diretto con altri modelli.
Domanda 9: Non esplicitamente menzionato l'uso previsto.
Domanda 10: Non si menzionano pregi o difetti espliciti, ma si capisce che il caricatore è un vantaggio.

Clip: clip_01.mp4
Domanda 1: Un uomo tiene un fucile in mano e parla rivolgendosi al pubblico, ringraziando e promuovendo una maglietta del canale.
Domanda 2: Informazione non disponibile.
Domanda 3: Informazione non disponibile.
Domanda 4: Informazione non disponibile.
Domanda 5: Informazione non disponibile.
Domanda 6: Informazione non disponibile.
Domanda 7: Informazione non disponibile.
Domanda 8: Informazione non disponibile.
Domanda 9: Informazione non disponibile.
Domanda 10: Informazione non disponibile.

Clip: clip_02.mp4
Domanda 1: Un uomo parla di un fucile Benelli, definendolo un modello leggendario e affidabile, mentre è sparato con munizioni military buck shot (doppio zero).
Domanda 2: Viene menzionato un fucile Benelli, ma il modello esatto non è specificato.
Domanda 3: Viene detto che è un design legacy per Benelli, indicato come "vacchiaro". Viene menzionata la munizione military buck shot doppio zero.
Domanda 4: Sì, viene mostrato l'uso pratico sparando con munizioni specifiche (military buck shot).
Domanda 5: Commento positivo, considerato un fucile fantastico e affidabile, con una reputazione duratura.
Domanda 6: Informazione non disponibile.
Domanda 7: Sì, sparare pratico con munizioni specifiche, ma risultati precisi non descritti.
Domanda 8: No confronto esplicito con altri modelli.
Domanda 9: Indicato come "vacchiaro", quindi probabilmente uso pratico generico, forse caccia o difesa (implicito).
Domanda 10: Pregi: affidabilità e design duraturo. Difetti non menzionati.
```

Figure 8: Esempio di risposte generate da Whisper+BLIP per le prime tre clip del video 6.

Infine, le risposte per tutte le clip sono state sintetizzate in un'unica sezione finale riassuntiva, costruita manualmente o tramite LLM a partire dalle risposte precedenti. La Figura 9 mostra un estratto di tale sintesi.

```
Clip: Totale Video
Domanda 1: Nel complesso, il video mostra un uomo che presenta e discute diversi aspetti di un fucile modello M2 (Benelli), parlando del design, accessori come caricatori a tamburo e chucks, spari pratici con munizioni diverse, uso tattico, sportivo e per forze dell'ordine.
Domanda 2: Il modello principale discusso è il fucile M2 di Benelli.
Domanda 3: Sono descritti dettagli tecnici sulla il caricatore a tamburo da 30 colpi, il calcio comfort tech rispetto al calcio distal grip, chuck evertibili e munizioni (military buck shot, slug, flights control).
Domanda 4: Sì, sono mostrati spari pratici con diverse munizioni e test su bersagli (anguria), anche se non sempre dettagliati.
Domanda 5: I commenti sono prevalentemente positivi, sottolineando affidabilità, leggerezza, bilanciamento e versatilità.
Domanda 6: Vengono menzionati caricatori a tamburo, chuck modificati, slugs, e calibri diversi come accessori o modifiche.
Domanda 7: Ci sono test di tiro su bersagli naturali (anguria) e uso pratico con munizioni specifiche.
Domanda 8: Sono fatti alcuni confronti tra calibri diversi (distal grip vs comfort tech), ma non tra modelli diversi di fucili.
Domanda 9: L'uso previsto include attività sportive (long gun), polizia/forze dell'ordine e uso generico affidabile.
Domanda 10: Vengono evidenziati pregi come affidabilità, design duraturo, leggerezza e bilanciamento; difetti riguardano l'idea di non modificare troppo il fucile e la preferenza per certe configurazioni (calibro dritto).
```

Figure 9: Sintesi finale delle informazioni ottenute da tutte le clip, generata a partire dalle risposte raccolte.

## 6 Confronto tra strategie e modelli

### 6.1 Confronto: Clip salienti vs. Video completo

Uno degli obiettivi centrali del nostro studio era valutare se l'approccio basato su **clip salienti**, ottenute tramite UniVTG, fosse preferibile rispetto all'analisi dell'intero video in un'unica soluzione.

A tal fine, abbiamo confrontato le risposte ottenute da Video-LLaVA su un video analizzato per intero (Video 2) con quelle generate analizzando segmenti brevi e rappresentativi dello stesso contenuto. Come mostrato negli screenshot in Figura 10 e Figura 11, emergono differenze significative nella qualità e nella granularità dell'informazione.

```

1. Cosa succede in questo clip?
In generale, nelle varie clip si vedono uomini (spesso soldati o appassionate) in uniforme o in abiti civili che presentano, mostrano, caricano e sparano diversi fucili, spesso alternando dimostrazioni pratiche (caricamento, sparare, test di tiro) a spiegazioni dettagliate delle caratteristiche dell'arma. Alcuni video includono anche confronti con altri modelli, opinioni personali, prove su bersagli fissi o mobili e dettagli di accessori.

2. Che modello di fucile viene mostrato o recensito?
I modelli principali sono:
M1 e M14: fucili d'assalto standard delle forze armate statunitensi, mostrati in più clip come principali oggetti di recensione.
M14: mostrato in una clip, evidenziato per la sua storia in conflitti e la potenza.
Fucile a pompa: presentato in almeno una clip, con dettagli sul funzionamento.
Fucile di precisione non specificato: mostrato in più video, a volte senza nome preciso.
In alcune clip si parla genericamente di "un fucile" senza modello chiaro.

3. Viene descritto qualche dettaglio tecnico del fucile?
Sì, vengono citati diversi dettagli tecnici:
Calibro di munizione (es. 5,56mm NATO per l'M14, 7,62 mm per l'M14).
Sistemi di munizione intercambiabili, a pompa o a fuoco singolo.
Sistemi di mira: ottiche, mirini ottici o a gas.
Meccanismi interni: spiegati in alcuni casi (caricatore, sincronizzazione del fuoco).
Aspetti del design: come mirini, silenziatori o configurazioni personalizzate.

4. Viene mostrato il funzionamento pratico del fucile?
Sì, quasi tutte le clip mostrano caricamento, sparare e talvolta dettagli del meccanismo interno. Alcuni video evidenziano il fucile in azione contro bersagli fissi o mobili, spesso come dimostrazione di precisione e potenza.

5. Ci sono commenti o opinioni sulle prestazioni del fucile?
Sì, in diversi casi si parla di:
Precisione elevata
Potenza di fuoco
Affidabilità, facilità di manutenzione.
Alcuni video menzionano anche aspetti negativi, come peso e ingombro.

6. Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?
Sì, molti video mostrano:
Silenziatori
Sistemi di mira
Sistemi di munizioni alternativi

```

Figure 10: Risposte generate da Video-LLaVA analizzando le clip salienti estratte dal Video 2.

```

clip_ambito_2.mp4: 1
"Cosa succede in questo clip?": "In questo clip, un uomo in uniforme militare spara un colpo con una pistola e poi si sposta verso il lato destro della schermata."
"Che modello di fucile viene mostrato o recensito?": "Il fucile viene mostrato e recensito in dettaglio, ma non viene specificato il modello."
"Viene descritto qualche dettaglio tecnico del fucile?": "Sì, il fucile viene descritto come dotato di un sistema di sincronizzazione."
"Viene mostrato il funzionamento pratico del fucile?": "Sì, il fucile viene mostrato in azione, con il protagonista che spara con precisione."
"Ci sono commenti o opinioni sulle prestazioni del fucile?": "Sì, il protagonista esprime la sua soddisfazione per le prestazioni del fucile, ma non viene specificato il motivo per cui è soddisfatto."
"Sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile?": "No, non sono menzionati o mostrati accessori, modifiche o personalizzazioni del fucile."
"Vengono mostrati test di tiro o prove pratiche?": "No, non sono mostrati test di tiro o prove pratiche."
"Viene fatto un confronto con altri modelli di fucile o armi simili?": "No, non viene fatto alcun confronto con altri modelli di fucile o armi simili."
"Si sono menzionati o utilizzati i nomi dei fucili?": "Sì, alcuni nomi sono menzionati, come M14, M16, M1, M2, M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, M13, M14, M15, M16, M17, M18, M19, M20, M21, M22, M23, M24, M25, M26, M27, M28, M29, M30, M31, M32, M33, M34, M35, M36, M37, M38, M39, M40, M41, M42, M43, M44, M45, M46, M47, M48, M49, M50, M51, M52, M53, M54, M55, M56, M57, M58, M59, M60, M61, M62, M63, M64, M65, M66, M67, M68, M69, M70, M71, M72, M73, M74, M75, M76, M77, M78, M79, M80, M81, M82, M83, M84, M85, M86, M87, M88, M89, M90, M91, M92, M93, M94, M95, M96, M97, M98, M99, M100, M101, M102, M103, M104, M105, M106, M107, M108, M109, M110, M111, M112, M113, M114, M115, M116, M117, M118, M119, M120, M121, M122, M123, M124, M125, M126, M127, M128, M129, M130, M131, M132, M133, M134, M135, M136, M137, M138, M139, M140, M141, M142, M143, M144, M145, M146, M147, M148, M149, M150, M151, M152, M153, M154, M155, M156, M157, M158, M159, M160, M161, M162, M163, M164, M165, M166, M167, M168, M169, M170, M171, M172, M173, M174, M175, M176, M177, M178, M179, M180, M181, M182, M183, M184, M185, M186, M187, M188, M189, M190, M191, M192, M193, M194, M195, M196, M197, M198, M199, M200, M201, M202, M203, M204, M205, M206, M207, M208, M209, M210, M211, M212, M213, M214, M215, M216, M217, M218, M219, M220, M221, M222, M223, M224, M225, M226, M227, M228, M229, M230, M231, M232, M233, M234, M235, M236, M237, M238, M239, M240, M241, M242, M243, M244, M245, M246, M247, M248, M249, M250, M251, M252, M253, M254, M255, M256, M257, M258, M259, M260, M261, M262, M263, M264, M265, M266, M267, M268, M269, M270, M271, M272, M273, M274, M275, M276, M277, M278, M279, M280, M281, M282, M283, M284, M285, M286, M287, M288, M289, M290, M291, M292, M293, M294, M295, M296, M297, M298, M299, M300, M301, M302, M303, M304, M305, M306, M307, M308, M309, M310, M311, M312, M313, M314, M315, M316, M317, M318, M319, M320, M321, M322, M323, M324, M325, M326, M327, M328, M329, M330, M331, M332, M333, M334, M335, M336, M337, M338, M339, M340, M341, M342, M343, M344, M345, M346, M347, M348, M349, M350, M351, M352, M353, M354, M355, M356, M357, M358, M359, M360, M361, M362, M363, M364, M365, M366, M367, M368, M369, M370, M371, M372, M373, M374, M375, M376, M377, M378, M379, M380, M381, M382, M383, M384, M385, M386, M387, M388, M389, M390, M391, M392, M393, M394, M395, M396, M397, M398, M399, M400, M401, M402, M403, M404, M405, M406, M407, M408, M409, M410, M411, M412, M413, M414, M415, M416, M417, M418, M419, M420, M421, M422, M423, M424, M425, M426, M427, M428, M429, M430, M431, M432, M433, M434, M435, M436, M437, M438, M439, M440, M441, M442, M443, M444, M445, M446, M447, M448, M449, M450, M451, M452, M453, M454, M455, M456, M457, M458, M459, M460, M461, M462, M463, M464, M465, M466, M467, M468, M469, M470, M471, M472, M473, M474, M475, M476, M477, M478, M479, M480, M481, M482, M483, M484, M485, M486, M487, M488, M489, M490, M491, M492, M493, M494, M495, M496, M497, M498, M499, M500, M501, M502, M503, M504, M505, M506, M507, M508, M509, M510, M511, M512, M513, M514, M515, M516, M517, M518, M519, M520, M521, M522, M523, M524, M525, M526, M527, M528, M529, M530, M531, M532, M533, M534, M535, M536, M537, M538, M539, M540, M541, M542, M543, M544, M545, M546, M547, M548, M549, M550, M551, M552, M553, M554, M555, M556, M557, M558, M559, M560, M561, M562, M563, M564, M565, M566, M567, M568, M569, M570, M571, M572, M573, M574, M575, M576, M577, M578, M579, M580, M581, M582, M583, M584, M585, M586, M587, M588, M589, M590, M591, M592, M593, M594, M595, M596, M597, M598, M599, M600, M601, M602, M603, M604, M605, M606, M607, M608, M609, M610, M611, M612, M613, M614, M615, M616, M617, M618, M619, M620, M621, M622, M623, M624, M625, M626, M627, M628, M629, M630, M631, M632, M633, M634, M635, M636, M637, M638, M639, M640, M641, M642, M643, M644, M645, M646, M647, M648, M649, M650, M651, M652, M653, M654, M655, M656, M657, M658, M659, M660, M661, M662, M663, M664, M665, M666, M667, M668, M669, M670, M671, M672, M673, M674, M675, M676, M677, M678, M679, M680, M681, M682, M683, M684, M685, M686, M687, M688, M689, M690, M691, M692, M693, M694, M695, M696, M697, M698, M699, M700, M701, M702, M703, M704, M705, M706, M707, M708, M709, M710, M711, M712, M713, M714, M715, M716, M717, M718, M719, M720, M721, M722, M723, M724, M725, M726, M727, M728, M729, M730, M731, M732, M733, M734, M735, M736, M737, M738, M739, M740, M741, M742, M743, M744, M745, M746, M747, M748, M749, M750, M751, M752, M753, M754, M755, M756, M757, M758, M759, M760, M761, M762, M763, M764, M765, M766, M767, M768, M769, M770, M771, M772, M773, M774, M775, M776, M777, M778, M779, M780, M781, M782, M783, M784, M785, M786, M787, M788, M789, M790, M791, M792, M793, M794, M795, M796, M797, M798, M799, M800, M801, M802, M803, M804, M805, M806, M807, M808, M809, M810, M811, M812, M813, M814, M815, M816, M817, M818, M819, M820, M821, M822, M823, M824, M825, M826, M827, M828, M829, M830, M831, M832, M833, M834, M835, M836, M837, M838, M839, M840, M841, M842, M843, M844, M845, M846, M847, M848, M849, M850, M851, M852, M853, M854, M855, M856, M857, M858, M859, M860, M861, M862, M863, M864, M865, M866, M867, M868, M869, M870, M871, M872, M873, M874, M875, M876, M877, M878, M879, M880, M881, M882, M883, M884, M885, M886, M887, M888, M889, M890, M891, M892, M893, M894, M895, M896, M897, M898, M899, M900, M901, M902, M903, M904, M905, M906, M907, M908, M909, M910, M911, M912, M913, M914, M915, M916, M917, M918, M919, M920, M921, M922, M923, M924, M925, M926, M927, M928, M929, M930, M931, M932, M933, M934, M935, M936, M937, M938, M939, M940, M941, M942, M943, M944, M945, M946, M947, M948, M949, M950, M951, M952, M953, M954, M955, M956, M957, M958, M959, M960, M961, M962, M963, M964, M965, M966, M967, M968, M969, M970, M971, M972, M973, M974, M975, M976, M977, M978, M979, M980, M981, M982, M983, M984, M985, M986, M987, M988, M989, M990, M991, M992, M993, M994, M995, M996, M997, M998, M999, M1000, M1001, M1002, M1003, M1004, M1005, M1006, M1007, M1008, M1009, M1010, M1011, M1012, M1013, M1014, M1015, M1016, M1017, M1018, M1019, M1020, M1021, M1022, M1023, M1024, M1025, M1026, M1027, M1028, M1029, M1030, M1031, M1032, M1033, M1034, M1035, M1036, M1037, M1038, M1039, M1040, M1041, M1042, M1043, M1044, M1045, M1046, M1047, M1048, M1049, M1050, M1051, M1052, M1053, M1054, M1055, M1056, M1057, M1058, M1059, M1060, M1061, M1062, M1063, M1064, M1065, M1066, M1067, M1068, M1069, M1070, M1071, M1072, M1073, M1074, M1075, M1076, M1077, M1078, M1079, M1080, M1081, M1082, M1083, M1084, M1085, M1086, M1087, M1088, M1089, M1090, M1091, M1092, M1093, M1094, M1095, M1096, M1097, M1098, M1099, M1100, M1101, M1102, M1103, M1104, M1105, M1106, M1107, M1108, M1109, M1110, M1111, M1112, M1113, M1114, M1115, M1116, M1117, M1118, M1119, M1120, M1121, M1122, M1123, M1124, M1125, M1126, M1127, M1128, M1129, M1130, M1131, M1132, M1133, M1134, M1135, M1136, M1137, M1138, M1139, M1140, M1141, M1142, M1143, M1144, M1145, M1146, M1147, M1148, M1149, M1150, M1151, M1152, M1153, M1154, M1155, M1156, M1157, M1158, M1159, M1160, M1161, M1162, M1163, M1164, M1165, M1166, M1167, M1168, M1169, M1170, M1171, M1172, M1173, M1174, M1175, M1176, M1177, M1178, M1179, M1180, M1181, M1182, M1183, M1184, M1185, M1186, M1187, M1188, M1189, M1190, M1191, M1192, M1193, M1194, M1195, M1196, M1197, M1198, M1199, M1200, M1201, M1202, M1203, M1204, M1205, M1206, M1207, M1208, M1209, M1210, M1211, M1212, M1213, M1214, M1215, M1216, M1217, M1218, M1219, M1220, M1221, M1222, M1223, M1224, M1225, M1226, M1227, M1228, M1229, M1230, M1231, M1232, M1233, M1234, M1235, M1236, M1237, M1238, M1239, M1240, M1241, M1242, M1243, M1244, M1245, M1246, M1247, M1248, M1249, M1250, M1251, M1252, M1253, M1254, M1255, M1256, M1257, M1258, M1259, M1260, M1261, M1262, M1263, M1264, M1265, M1266, M1267, M1268, M1269, M1270, M1271, M1272, M1273, M1274, M1275, M1276, M1277, M1278, M1279, M1280, M1281, M1282, M1283, M1284, M1285, M1286, M1287, M1288, M1289, M1290, M1291, M1292, M1293, M1294, M1295, M1296, M1297, M1298, M1299, M1300, M1301, M1302, M1303, M1304, M1305, M1306, M1307, M1308, M1309, M1310, M1311, M1312, M1313, M1314, M1315, M1316, M1317, M1318, M1319, M1320, M1321, M1322, M1323, M1324, M1325, M1326, M1327, M1328, M1329, M1330, M1331, M1332, M1333, M1334, M1335, M1336, M1337, M1338, M1339, M1340, M1341, M1342, M1343, M1344, M1345, M1346, M1347, M1348, M1349, M1350, M1351, M1352, M1353, M1354, M1355, M1356, M1357, M1358, M1359, M1360, M1361, M1362, M1363, M1364, M1365, M1366, M1367, M1368, M1369, M1370, M1371, M1372, M1373, M1374, M1375, M1376, M1377, M1378, M1379, M1380, M1381, M1382, M1383, M1384, M1385, M1386, M1387, M1388, M1389, M1390, M1391, M1392, M1393, M1394, M1395, M1396, M1397, M1398, M1399, M1400, M1401, M1402, M1403, M1404, M1405, M1406, M1407, M1408, M1409, M1410, M1411, M1412, M1413, M1414, M1415, M1416, M1417, M1418, M1419, M1420, M1421, M1422, M1423, M1424, M1425, M1426, M1427, M1428, M1429, M1430, M1431, M1432, M1433, M1434, M1435, M1436, M1437, M1438, M1439, M1440, M1441, M1442, M1443, M1444, M1445, M1446, M1447, M1448, M1449, M1450, M1451, M1452, M1453, M1454, M1455, M1456, M1457, M1458, M1459, M1460, M1461, M1462, M1463, M1464, M1465, M1466, M1467, M1468, M1469, M1470, M1471, M1472, M1473, M1474, M1475, M1476, M1477, M1478, M1479, M1480, M1481, M1482, M1483, M1484, M1485, M1486, M1487, M1488, M1489, M1490, M1491, M1492, M1493, M1494, M1495, M1496, M1497, M1498, M1499, M1500, M1501, M1502, M1503, M1504, M1505, M1506, M1507, M1508, M1509, M1510, M1511, M1512, M1513, M1514, M1515, M1516, M1517, M1518, M1519, M1520, M1521, M1522, M1523, M1524, M1525, M1526, M1527, M1528, M1529, M1530, M1531, M1532, M1533, M1534, M1535, M1536, M1537, M1538, M1539, M1540, M1541, M1542, M1543, M1544, M1545, M1546, M1547, M1548, M1549, M1550, M1551, M1552, M1553, M1554, M1555, M1556, M1557, M1558, M1559, M1560, M1561, M1562, M1563, M1564, M1565, M1566, M1567, M1568, M1569, M1570, M1571, M1572, M1573, M1574, M1575, M1576, M1577, M1578, M1579, M1580, M1581, M1582, M1583, M1584, M1585, M1586, M1587, M1588, M1589, M1590, M1591, M1592, M1593, M1594, M1595, M1596, M1597, M1598, M1599, M1600, M1601, M1602, M1603, M1604, M1605, M1606, M1607, M1608, M1609, M1610, M1611, M1612, M1613, M1614, M1615, M1616, M1617, M1618, M1619, M1620, M1621, M1622, M1623, M1624, M1625, M1626, M1627, M1628, M1629, M1630, M1631, M1632, M1633, M1634, M1635, M1636, M1637, M1638, M1639, M1640, M1641, M1642, M1643, M1644, M1645, M1646, M1647, M1648, M1649, M1650, M1651, M1652, M1653, M1654, M1655, M1656, M1657, M1658, M1659, M1660, M1661, M1662, M1663, M1664, M1665, M1666, M1667, M1668, M1669, M1670, M1671, M1672, M1673, M1674, M1675, M1676, M1677, M1678, M1679, M1680, M1681, M1682, M1683, M1684, M1685, M1686, M1687, M1688, M1689, M1690, M1691, M1692, M1693, M1694, M1695, M1696, M1697, M1698, M1699, M1700, M1701, M1702, M1703, M1704, M1705, M1706, M1707, M1708, M1709, M1710, M1711, M1712, M1713, M1714, M1715, M1716, M1717, M1718, M1719, M1720, M1721, M1722, M1723, M1724, M1725, M1726, M1727, M1728, M1729, M1730, M1731, M1732, M1733, M1734, M1735, M1736, M1737, M1738, M1739, M1740, M1741, M1742, M1743, M1744, M1745, M1746, M1747, M1748, M1749, M1750, M1751, M1752, M1753, M1754, M1755, M1756, M1757, M1758, M1759, M1760, M1761, M1762, M1763, M1764, M1765, M1766, M1767, M1768, M1769, M1770, M1771, M1772, M1773, M1774, M1775, M1776, M1777, M1778, M1779, M1780, M1781, M1782, M1783, M1784, M1785, M1786, M1787, M1788, M1789, M1790, M1791, M1792, M1793, M1794, M1795, M1796, M1797, M1798, M1799, M1800, M1801, M1802, M1803, M1804, M1805, M1806, M1807, M1808, M1809, M1810, M1811, M1812, M1813, M1814, M1815, M1816, M1817, M1818, M1819, M1820, M1821, M1822, M1823, M1824, M1825, M1826, M1827, M1828, M1829, M1830, M1831, M1832, M1833, M1834, M1835, M1836, M1837, M1838, M1839, M1840, M1841, M1842, M1843, M1844, M1845, M1846, M1847, M1848, M1849, M1850, M1851, M1852, M1853, M1854, M1855, M1856, M1857, M1858, M1859, M1860, M1861, M1862, M1863, M1864, M1865, M1866, M1867, M1868, M1869, M1870, M1871, M1872, M1873, M1874, M1875, M1876, M1877, M1878, M1879, M1880, M1881, M1882, M1883, M1884, M1885, M1886, M1887, M1888, M1889, M1890, M1891, M1892, M1893, M1894, M1895, M1896, M1897, M1898, M1899, M1900, M1901, M1902, M1903, M1904, M1905, M1906, M1907, M1908, M1909, M1910, M1911, M1912, M1913, M1914, M1915, M1916, M1917, M1918, M1919, M1920, M1921, M1922, M1923, M1924, M1925, M1926, M1927, M1928, M1929, M1930, M1931, M1932, M1933, M1934, M1935, M1936, M1937, M1938, M1939, M1940, M1941, M1942, M1943, M1944, M1945, M1946, M1947, M1948, M1949, M1950, M1951, M1952, M1953, M1954, M1955, M1956, M1957, M1958, M1959, M1960, M1961, M1962, M1963, M1964, M1965, M1966, M1967, M1968, M1969, M1970, M1971, M1972, M1973, M1974, M1975, M1976, M1977, M1978, M1979, M1980, M1981, M1982, M1983, M1984, M1985, M1986, M1987, M1988, M1989, M1990, M1991, M1992, M1993, M1994, M1995, M1996, M1997, M1998, M1999, M2000, M2001, M2002, M2003, M2004, M2005, M2006, M2007, M2008, M2009, M2010, M2011, M2012, M2013, M2014, M2015, M2016, M2017, M2018, M2019, M2020, M2021, M2022, M2023, M2024, M2025, M2026, M2027, M2028, M2029, M2030, M2031, M2032, M2033, M2034, M2035, M2036, M2037, M2038, M2039, M2040, M2041, M2042, M2043, M2044, M2045, M2046, M2047, M2048, M2049, M2050, M2051, M2052, M2053, M2054, M2055, M2056, M2057, M2058, M2059, M2060, M2061, M2062, M2063, M2064, M2065, M2066, M2067, M2068, M2069, M2070, M2071, M2072, M2073, M2074, M2075, M20
```

## 6.2 Video-LLaVA vs. Whisper+BLIP

### 6.3 Confronto tra strategie multimodali: Video-LLaVA vs. Whisper+BLIP

Il secondo confronto ha riguardato le due strategie multimodali adottate: Video-LLaVA da un lato, e l'approccio modulare Whisper+BLIP dall'altro.

Whisper+BLIP si è rivelato particolarmente efficace per l'analisi di video con forte componente **audio descrittiva**, ottenendo un'accuracy superiore nella corrispondenza con il ground truth. Questo è emerso soprattutto nei casi in cui il contenuto parlato copriva gran parte delle informazioni tecniche e contestuali. I risultati sono sintetizzati nella Figura 12, che riporta una tabella comparativa delle risposte corrette generate dai due sistemi per ciascun video.

	VideoLLava	Blip+Whisper
Video 1	5,8,9,10	2,3,4,5,7,8,9,10
Video 2	1,3,4,6,7,9,10	1,2,5,7,8,9,10
Video 3	1,4,5,9,10	1,2,5,6,7,8,9,10
Video 4	3,5,9,10	1,3,4,5,6,7,8,10
Video 5	1,3,4,7,9	2,4,8,9
Video 6	1,4,7,9,10	1,2,4,7,8,9,10
Video 8	1,4,6,10	1,2,3,5,6,7,8,9,10
Video 9	3,4,6,9	3,5,6,7,8,9,10
Video 10	3,4,8,9	1,2,3,4,6,10
TOTALE CORRETTE	42	64
ACCURACY	47%	71%

Figure 12: Accuracy delle risposte fornite da Video-LLaVA e Whisper+BLIP rispetto al ground truth. I valori indicano il numero di risposte corrette su dieci domande.

Tuttavia, questa maggiore accuratezza si limita spesso a contenuti esplicitamente pronunciati: l'approccio Whisper+BLIP risulta meno efficace quando l'informazione è veicolata da elementi visivi, come accessori, gesti, ambientazioni o funzionalità mostrate in azione.

In questo contesto, Video-LLaVA ha evidenziato una superiore **capacità di interpretazione visiva**, riuscendo a cogliere dettagli che Whisper+BLIP ignora. Un esempio emblematico è il **Video 5**, in cui Video-LLaVA ha identificato la modalità convertibile del Benelli M3 (pump/semi-auto), la presenza di un calcio telescopico e altre caratteristiche strutturali, elementi non rilevati da Whisper. Tali dettagli erano chiaramente riconoscibili nelle immagini, ma non verbalizzati dal soggetto parlante.

Le Figure 13 e 14 mostrano il confronto qualitativo tra i due approcci. Mentre Whisper+BLIP fornisce descrizioni accurate del parlato, Video-LLaVA restituisce una visione più completa e contestuale della scena. Questo conferma

l'utilità di approcci visivo-linguistici completi per l'analisi semantica di video ad alto contenuto visivo.

1. Cosa succede nel video?  
Il video mostra diverse scene in cui un uomo (o più uomini) presenta, mostra e utilizza varie armi da fuoco, tra cui fucili e pistole. Si vedono azioni di sparò sia in ambienti chiusi (capanna, panca) sia in esterni (campo di tiro, pista di tiro). Vengono mostrati il caricamento, il meccanismo di funzionamento e l'uso pratico delle armi, con alcuni test di tiro su bersagli diversi.

Figure 13: Risposte generate da Video-LLaVA per le clip del Video 5.

Clip: Totale Video  
Domanda 1: Nel video complessivo viene mostrata una recensione dettagliata di fucili Beretta, con focus particolare sul modello M4 e alcuni riferimenti agli M2 e M3. Vengono illustrati aspetti tecnici, funzionali e pratici dell'arma, insieme a scene di tiro e spiegazioni sulle regolazioni e funzionamento.

Figure 14: Trascrizione e descrizioni generate da Whisper+BLIP per il Video 5.

## 6.4 Conclusione del confronto

Il confronto evidenzia che:

- L'approccio basato su **clip salienti** è più efficace del video intero per l'estrazione di contenuti ricchi, tecnici e distribuiti nel tempo.
- Whisper+BLIP è adatto a scenari a bassa complessità visiva, ma con elevata densità informativa nella traccia audio.
- Video-LLaVA offre una comprensione multimodale completa, risultando più robusto quando l'audio è scarso o assente e le informazioni sono presenti solo nella componente visiva.

Questa analisi ha guidato la scelta finale di combinare UniVTG e Video-LLaVA come pipeline principale per la comprensione automatica dei video.

## 7 Conclusioni e sviluppi futuri

### Conclusioni

Il progetto ha affrontato il problema della comprensione automatica di contenuti video tramite modelli di intelligenza artificiale multimodale, con l'obiettivo di costruire una pipeline in grado di estrarre descrizioni testuali significative da contenuti audiovisivi complessi.

La soluzione sviluppata ha integrato il modello UniVTG per la selezione automatica dei momenti salienti, e Video-LLaVA per la generazione contestuale di risposte a domande specifiche, affiancati da una pipeline alternativa basata su Whisper e BLIP in scenari a ridotta disponibilità computazionale.

Durante le fasi iniziali di sviluppo, l'intero sistema è stato eseguito su *Google Colab*. Tuttavia, le limitazioni di memoria e instabilità riscontrate con modelli di larga scala come Video-LLaVA hanno reso necessario migrare l'esecuzione su infrastrutture locali più potenti. In particolare, le fasi di inferenza sono state portate su una **GPU NVIDIA A6000** (48 GB di VRAM) messa a disposizione dall'ateneo, in grado di ospitare i modelli completi e garantire maggiore affidabilità durante l'esecuzione.

Nonostante alcune difficoltà tecniche legate alla gestione delle dipendenze e alla scalabilità, il sistema è risultato funzionale: la pipeline è operativa, le clip vengono generate in modo coerente, e le risposte testuali mostrano una buona capacità descrittiva, soprattutto nei casi in cui è disponibile una forte componente visiva.

Le limitazioni principali non sono da attribuirsi a scelte progettuali errate, ma piuttosto alla complessità attuale degli strumenti di deep learning in ambito multimodale e all'elevato carico computazionale richiesto. L'esperienza maturata ha comunque permesso di validare il flusso di lavoro e gettare le basi per sviluppi futuri più robusti.

### Sviluppi futuri

L'esperienza condotta ha evidenziato numerose aree di miglioramento che possono costituire la base per sviluppi successivi.

Un primo obiettivo riguarda il consolidamento del deployment del sistema, attraverso l'adozione di ambienti persistenti (come container Docker o server universitari), al fine di garantire stabilità nelle dipendenze e continuità nelle esecuzioni. In parallelo, sarà importante ridurre la dipendenza da piattaforme come Google Colab, orientandosi verso soluzioni cloud o locali più scalabili.

Sul piano linguistico, è possibile affinare la qualità delle descrizioni generate intervenendo sulla progettazione dei prompt o introducendo tecniche di fine-tuning su dataset specifici del dominio. L'integrazione di strategie di prompt chaining o question re-ranking potrebbe inoltre aumentare la coerenza semantica delle risposte prodotte.

Infine, una possibile estensione del progetto prevede lo sviluppo di un'interfaccia interattiva che permetta l'utilizzo del sistema da parte di utenti non esperti, con l'obiettivo di rendere accessibile l'intero flusso di analisi e sintesi video. Questo traguardo richiederà la standardizzazione delle pipeline e la validazione su un più ampio insieme di dati reali.

Il valore principale del lavoro risiede nella sua funzione esplorativa e nella capacità di individuare con chiarezza le criticità ancora aperte nell'ambito dell'intelligenza artificiale multimodale, contribuendo così alla riflessione su sistemi scalabili, affidabili e adattabili per la sintesi automatica di contenuti audiovisivi.

## References

- [1] Glenn Jocher et al. *YOLOv5*. Ultralytics implementation. 2020. URL: <https://github.com/ultralytics/yolov5>.
- [2] Alexei Baevski et al. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Advances in Neural Information Processing Systems* (2020).
- [3] Zhongzheng Chen, Kevin Lin, Zhicheng Yu, et al. “UniVTG: Unified Video-Text Grounding”. In: *arXiv preprint arXiv:2306.01792* (2023). URL: <https://github.com/showlab/UniVTG>.
- [4] Awni Hannun et al. “Deep Speech: Scaling up end-to-end speech recognition”. In: *arXiv preprint arXiv:1412.5567* (2014).
- [5] Kaiming He et al. “Mask R-CNN”. In: *IEEE International Conference on Computer Vision (ICCV)* (2017).
- [6] Junnan Li et al. “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation”. In: *arXiv preprint arXiv:2201.12086* (2022). URL: <https://github.com/salesforce/BLIP>.
- [7] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. “Unsupervised video summarization with adversarial LSTM networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
- [8] Danila Potapov et al. “Category-specific video summarization”. In: *European Conference on Computer Vision (ECCV)* (2014).
- [9] Alec Radford et al. “Whisper: Robust Speech Recognition via Large-Scale Weak Supervision”. In: *OpenAI* (2022). URL: <https://github.com/openai/whisper>.
- [10] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Advances in neural information processing systems* 28 (2015).
- [11] Mingxing Tan and Quoc V. Le. “EfficientDet: Scalable and Efficient Object Detection”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
- [12] LanguageBind team. “Video-LLaVA-7B”. In: (2024). Model weights. URL: <https://huggingface.co/LanguageBind/Video-LLaVA-7B>.
- [13] Junyang Wang, Yixiao He, Zekai Zhang, et al. “Video-LLaVA: Learning Multimodal LLMs for Video Understanding”. In: *arXiv preprint arXiv:2403.10660* (2024). URL: <https://github.com/PKU-YuanGroup/Video-LLaVA>.
- [14] Ke Zhou, Yueming Qiao, and Tao Xiang. “Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward”. In: *AAAI Conference on Artificial Intelligence* (2018).