



UNIVERSITÀ DEGLI STUDI DI CATANIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA
CORSO DI LAUREA TRIENNALE IN INFORMATICA

Kevin Speranza

Egocentric Videos Understanding with Active Memory

RELAZIONE PROGETTO FINALE

Relatore: Francesco Ragusa

Anno Accademico 2024 - 2025

Abstract

L'abstract va inserito qui.

Indice

1	Introduzione	3
2	Lavori correlati	6
2.1	Long video understanding	6
2.2	Structured video representations	7
2.3	Video summarization	7
3	Metodo - AMEGO	9
3.1	Decomposizione del video	9
3.2	Costruzione della memoria	10
3.2.1	Object interaction tracklets	10
3.2.2	Location segments	14
3.3	Pseudocode	15
3.3.1	Object interactions	16
3.3.2	Location Segment	17
3.4	AMB - Active Memories Benchmark	18
3.4.1	Tipologie di query	18
3.4.2	Risposte alle Query	19
3.5	Risultati	19
4	Dataset	21
5	Esperimenti	22
6	Risultati	23
	Conclusione	24
	Bibliografia	25

Capitolo 1

Introduzione

Negli ultimi anni i dispositivi indossabili per la registrazione di video in prima persona hanno conosciuto una diffusione sempre più ampia. Strumenti come *smart glasses*, *body cameras*, *action cameras* hanno reso possibile la cattura di flussi visivi continui dal punto di vista diretto dell'utilizzatore, dando origine a quella che viene comunemente definita come *egocentric video*. Questa modalità di acquisizione ha suscitato un forte interesse [1] non solo per le applicazioni pratiche, che spaziano dall'intrattenimento personale alla sicurezza e al monitoraggio di attività lavorative, ma anche per le sfide che pone in termini di analisi e interpretazione dei contenuti.



Figura 1.1: Esempi di dispositivi indossabili per la cattura di video egocentrici

L'adozione di tali dispositivi è stata favorita dalla loro versatilità: da un lato vengono utilizzati per scopi ricreativi e per la condivisione di esperienze personali, dall'altro trovano applicazione in contesti professionali e industriali, dove consentono di documentare procedure complesse e migliorare i processi di formazione e supervisione. Ciò che rende peculiari i video egocentrici è la loro capacità di catturare dettagli e prospettive uniche, fornendo una visione diretta dell'attività di chi li indossa.

Il principale ostacolo all'analisi di questi contenuti risiede nella loro natura non strutturata. I video in prima persona possono essere considerati

come veri e propri flussi di coscienza visivi: lunghi, frammentati, privi di un’organizzazione narrativa chiara e difficili da interpretare. La presenza di movimenti rapidi della fotocamera, variazioni di illuminazione e interazioni simultanee con più oggetti rende complicata l’estrazione di significato. Un’annotazione manuale completa non è praticabile, sia per la mole di dati prodotta sia per la complessità dei contenuti.

Da questa problematica emerge la necessità di costruire una “memoria artificiale” capace di trasformare i video egocentrici in rappresentazioni strutturate e interrogabili. Questa tesi prende avvio dall’analisi dei principali contributi degli elementi che vanno a formare un eventuale memoria artificiale, per poi focalizzarsi su AMEGO [2], acronimo di *Active Memory of the EGOCentric video*, un sistema sviluppato per organizzare e rendere interrogabili i contenuti visivi in prima persona e attualmente considerato stato dell’arte nel suo ambito [2].

Il contributo di questo lavoro consiste nella valutazione di AMEGO in un contesto diverso rispetto a quello in cui è stato originariamente validato. Inizialmente il sistema è stato sperimentato sul dataset *EPIC KITCHENS* [3], una collezione di video ambientati in cucine domestiche. In questa tesi, invece, viene preso in esame il dataset *ENIGMA-51* [4], un dataset egocentrico acquisito in scenari industriali, in cui diversi operatori hanno seguito procedure guidate per eseguire attività di riparazione di quadri elettrici. La differenza tra i due domini rende lo studio particolarmente interessante, in quanto consente di valutare la capacità di AMEGO di generalizzare a contesti applicativi mai visti.

In ambito industriale, l’analisi dei video egocentrici e la gestione della concurrency assumono un ruolo fondamentale per ottimizzare diversi aspetti operativi. In particolare, si possono individuare due benefici principali:

- **Affidabilità del processo:** garantire che le operazioni vengano eseguite nell’ordine corretto consente di ottimizzare i flussi produttivi e di ridurre il rischio di errori umani, assicurando maggiore coerenza nelle procedure.
- **Sicurezza dei lavoratori:** monitorare in tempo reale l’uso corretto dei dispositivi di protezione individuale (DPI) oppure verificare la collocazione di componenti critici. Ad esempio, un materiale pericoloso come un condensatore ad alta tensione deve essere riposto in aree designate per prevenire incidenti.

Un aspetto centrale analizzato in questo lavoro riguarda la gestione della *concurrency*, intesa come la capacità di riconoscere non solo con quali altri oggetti un determinato strumento viene utilizzato in simultanea, ma anche in

quali contesti o aree operative tale oggetto viene impiegato. Nei capitoli successivi verranno prima esaminati gli strumenti e le metodologie attualmente presenti in letteratura, che costituiscono le basi per lo sviluppo di sistemi di memoria artificiale per video egocentrici. Successivamente saranno illustrate nel dettaglio le caratteristiche di AMEGO [2] e le sperimentazioni condotte sul dataset ENIGMA-51 [4], con l'obiettivo di valutare fino a che punto il sistema possa essere adattato a contesti applicativi diversi da quelli per cui è stato originariamente progettato.

Capitolo 2

Lavori correlati

2.1 Long video understanding

La comprensione di video di lunga durata riguarda l’analisi e l’interpretazione di contenuti visivi che possono estendersi per diversi minuti o addirittura ore. Questi video richiedono metodi capaci di catturare sequenze temporali complesse e le interazioni tra più oggetti e persone, il che rende la gestione di flussi video così lunghi particolarmente impegnativa dal punto di vista computazionale.

Per affrontare la sfida della comprensione di video di lunga durata, sono stati sviluppati benchmark specifici che mettono alla prova la capacità dei modelli di gestire sequenze temporali estese e interazioni complesse. Tra i più rilevanti troviamo EgoSchema [5], un dataset egocentrico con video della durata massima di circa tre minuti, progettato per diagnosticare azioni quotidiane articolate in più passaggi. I video includono azioni come preparare un oggetto, combinarlo con altri strumenti o spostarlo tra diverse aree operative, spesso con oggetti in movimento e interazioni parzialmente sovrapposte. I modelli devono quindi seguire la coerenza temporale, riconoscere pattern ricorrenti, e inferire correttamente le relazioni tra oggetti e azioni, mantenendo una rappresentazione accurata di sequenze multi-step.

ReST [6] propone invece scenari industriali e scientifici più complessi, con video più lunghi in cui interagiscono simultaneamente più strumenti, oggetti e operatori. I compiti richiesti includono il tracciamento di oggetti su intervalli temporali estesi, il riconoscimento di interazioni multiple e la comprensione delle relazioni spaziali tra componenti e strumenti.

Diversi approcci sono stati sviluppati per la comprensione di video di lunga durata. Alcuni trattano il problema come un task di *natural language question answering*, generando prima dei sottotitoli o descrizioni automatiche

del video e utilizzando LLM per rispondere a domande specifiche [7, 8, 9, 10, 11, 12]. Altri approcci integrano direttamente LLM con un encoder video, sfruttando le capacità di comprensione e generazione dei modelli linguistici per elaborare sequenze visive estese in maniera coerente [13, 14, 15, 16].

2.2 Structured video representations

Con *rappresentazione strutturata* si intende l’insieme di tecniche volte a organizzare un video non come una semplice sequenza di frame, ma come una struttura semantica in grado di esplicitare le relazioni tra gli elementi presenti nella scena. Questo approccio consente di passare da una descrizione puramente visiva a una rappresentazione schematica e strutturata, che rende possibile interrogare i video in maniera più efficace, permettendo di estrarre informazioni mirate.

Un filone centrale della ricerca si è concentrato sullo studio delle relazioni contestuali, investigando in particolare i legami tra oggetti e attori [17, 18, 19, 20, 21, 22, 23, 24]. Parallelamente, sono stati proposti modelli basati su grafi per rappresentare le dipendenze tra azioni, al fine di catturare la dimensione temporale e causale dei comportamenti.

Alcuni lavori hanno cercato di spingersi oltre, introducendo strategie più specifiche. Un esempio è UnweaveNet [25], che raggruppa i video in *activity threads*, ossia insiemi di clip collegati logicamente che consentono di separare e ricostruire le diverse “storie” di attività intrecciate all’interno di una sequenza più lunga. Un altro contributo rilevante è l’introduzione dei cosiddetti *egocentric scene graphs* [26], strutture orientate a rappresentare in modo esplicito le interazioni tra il soggetto che indossa la videocamera e gli oggetti presenti nell’ambiente.

Nonostante questi progressi, tali approcci risultano ancora limitati, poiché tendono a catturare soltanto alcuni aspetti delle attività. In particolare, faticano a integrare in un unico modello le molteplici dimensioni tipiche dei video egocentrici: le interazioni con gli oggetti, i luoghi chiave in cui esse avvengono e le interdipendenze tra questi elementi. Questa mancanza di completezza riduce la capacità di ottenere rappresentazioni realmente efficaci per la comprensione e la memorizzazione dei flussi visivi in prima persona.

2.3 Video summarization

Il riassunto del video ha come obiettivo la generazione di una versione ridotta di un video, tipicamente attraverso l’estrazione di *key frames* o *key shots* che

ne catturino i momenti salienti. Gli approcci proposti in letteratura variano in funzione degli elementi ritenuti rilevanti per la sintesi: alcuni si focalizzano sulla presenza e sul ruolo delle persone o degli oggetti all'interno della scena [27], altri privilegiano la rilevazione di eventi significativi [28], mentre ulteriori metodi considerano anche caratteristiche estetiche dei fotogrammi chiave per selezionare i contenuti più rappresentativi [29].

Accanto a questi, sono stati introdotti approcci in grado di generare i riassunti in modalità *online*, ossia durante la riproduzione del flusso video, consentendo una sintesi in tempo reale [30, 31]. Tuttavia, tali tecniche non costruiscono una rappresentazione strutturata del video e risultano spesso sensibili al rumore prodotto dai modelli di rilevamento, mancando di un'efficace integrazione della dimensione temporale.

Un contributo particolarmente rilevante in questa direzione è rappresentato da [32], che introduce una modalità di sintesi per i video egocentrici basata su attori, eventi, luoghi e oggetti. Questo approccio permette di interrogare i contenuti lungo diverse dimensioni, anche combinandole attraverso operatori booleani. Ciononostante, il metodo rimane in gran parte vincolato al riconoscimento di attrazioni predefinite e di oggetti visivamente distinti, risultando meno efficace in scenari affollati e caotici.

Capitolo 3

Metodo - AMEGO

AMEGO, acronimo di *Active Memory of the EGOCentric video*, è concepito per trasformare un video egocentrico lungo e non strutturato in una memoria capace di descrivere in modo completo le interazioni del soggetto con oggetti e luoghi, e al tempo stesso interrogabile per recuperare i segmenti temporali in cui un oggetto è stato utilizzato, una location visitata o entrambe le condizioni si sono verificate congiuntamente.

Un aspetto cruciale che distingue AMEGO da altri approcci riguarda la sua natura *semantic-free*. Gli oggetti e le location non vengono legati ad una tassonomia fissa di etichette o ad un vocabolario prestabilito. Essi vengono invece rappresentati direttamente sulla base delle caratteristiche visive, consentendo così una distinzione più fine e dettagliata tra le diverse istanze. Questo approccio permette al sistema di adattarsi a contesti nuovi senza la necessità di ridefinire un insieme di categorie predefinite.

3.1 Decomposizione del video

Dato un video egocentrico \mathcal{V} , esso viene scomposto in due elementi fondamentali:

- **Hand-Object Interaction (HOI) tracklets**, indicati con Θ : ciascun HOI tracklet¹ descrive in maniera spazio-temporale un oggetto che interagisce in modo consistente con almeno una mano del soggetto. Ogni

¹**Tracklet**: sequenza di bounding box che identifica in modo coerente la traiettoria o l'interazione di un oggetto nel tempo.

tracklet è caratterizzato da bounding boxes² spazio-temporali e dalle corrispondenti feature visive³.

- **Location segments**, indicati con \mathcal{L} : ogni elemento corrisponde a un intervallo temporale in cui il soggetto si trova in un determinato luogo e vi svolge interazioni. L'interesse è focalizzato sulle cosiddette *activity-centric-zones*, ossia i luoghi in cui avvengono le principali interazioni con gli oggetti.

Combinando gli *HOI tracklets* con i *Location segments* si ottiene una memoria strutturata in grado di eseguire i compiti discussi in precedenza.

3.2 Costruzione della memoria

La memoria viene definita come:

$$\mathcal{E} = \{\mathcal{O}, \mathcal{L}\}$$

dove:

- \mathcal{E} : AMEGO
- \mathcal{O} : insieme di HOI tracklets
- \mathcal{L} : insieme dei Location Segments.

Questa memoria viene costruita *online*, eliminando la necessità di riprocessare continuamente informazioni passate.

3.2.1 Object interaction tracklets

Gli *HOI tracklets*, indicati con \mathcal{O} rappresentano sequenze di interazioni tra le mani del soggetto e gli oggetti presenti nel video. Formalmente, possiamo definire l'insieme degli HOI tracklets come:

$$\mathcal{O} = \{o_1, o_2, \dots, o_n\}$$

dove ciascun tracklet $o_i \in \mathcal{O}$ è una tupla:

$$o_i = (t_s, t_e, b_t, h, \text{id})$$

con:

²**Bounding box**: regione rettangolare che delimita un oggetto in un singolo frame del video.

³**Feature visive**: rappresentazioni numeriche delle proprietà visive di un oggetto

- t_s : istante di inizio dell'interazione
- t_e : istante di fine dell'interazione
- b_t : sequenza di bounding box che raffigurano l'oggetto
- h : lato della mano che compie l'interazione (sinistra o destra)
- id: identificatore dell'istanza dell'oggetto associato al tracklet

La costruzione della memoria \mathcal{O} avviene in maniera iterativa, processando il video frame per frame tramite una pipeline composta da tre fasi principali:

1. **Initialisation:** individuazione dei possibili nuovi HOI tracklets.
2. **Updating:** aggiornamento dei tracklets attivi⁴, corrispondenti alle interazioni in corso.
3. **Assignment and storing:** i tracklets terminati⁵ vengono archiviati nella memoria \mathcal{E} e viene assegnata loro l'istanza oggetto corrispondente.

Inizializzazione

La prima fase consiste nell'individuazione dei nuovi *HOI tracklets*. Per questo utilizziamo un detector di *hand-object-interaction class-agnostic*⁶ [33], che fornisce insiemi di bounding box attive per oggetti e mani, denotati rispettivamente come \mathcal{B}_t^o e \mathcal{B}_t^h .

Un nuovo *HOI tracklet* o_i viene inizializzato per ciascuna nuova hand-object-interaction rilevata. Ogni tracklet è definito come una sequenza di almeno s_o bounding box che mostrano un forte sovrapposizione spaziale all'interno di una finestra temporale di w_s frame.

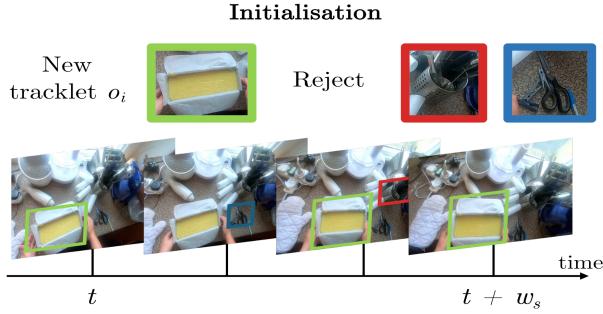
Questo filtraggio spazio-temporale permette di ridurre il rumore derivante dall'applicazione indipendente del rilevatore su ciascun frame. Considerando la durata naturale delle interazioni mano-oggetto, è possibile identificare in modo affidabile i nuovi tracklets attivi, garantendo coerenza spaziale e temporale nelle rilevazioni.

Il tracklet o_i viene ora considerato *attivo* e aggiunto alla memoria \mathcal{O} .

⁴**Tracklet attivi:** tracklets che stanno effettivamente registrando un'interazione in corso tra la mano del soggetto e l'oggetto

⁵**Tracklet terminato:** l'azione per cui veniva considerato attivo è terminata

⁶**class-agnostic detector:** non fa distinzione tra classi predefinite di oggetti, ma identifica interazioni tra mani e oggetti basandosi su caratteristiche visive generiche

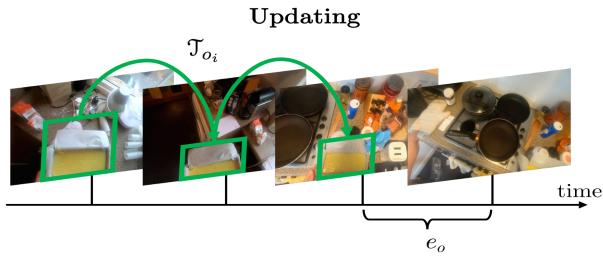
**Figura 3.1:** Fase di inizializzazione

Per ciascun frame successivo, calcoliamo l'*Intersection over Union (IoU)*⁷ tra i bounding box degli oggetti che interagiscono con la stessa mano. I bounding box che superano una soglia θ vengono assegnati al tracklet o_i . Se non è possibile assegnare nuovi bounding box al tracklet, questo viene considerato *terminato*.

Updating

Questa fase mira a catturare l'intera durata dell'interazione e contemporaneamente a seguire tutte le occorrenze spaziali dell'oggetto.

Sebbene i rilevatori di HOI a livello di singolo frame siano sufficienti per identificare nuove interazioni, essi non sono in grado di estendere in modo affidabile i tracklets quando mani o oggetti escono dal campo visivo egocentrico. Per questo motivo, viene utilizzato un off-the-shelf *single-object tracker (SOT)*⁸ [34].

**Figura 3.2:** Fase di updating

⁷**Intersection over Union (IoU):** misura di sovrapposizione tra due bounding box, calcolata come il rapporto tra l'area di intersezione e l'area di unione dei due rettangoli.

⁸**Single-Object Tracker (SOT):** permette di seguire un singolo oggetto nel tempo, stimandone la posizione frame per frame anche in assenza di rilevazioni dirette.

Per ogni tracklet attivo o_i viene inizializzato un SOT. Consideriamo il tracklet o_i terminato se non sono presenti rilevazioni associate \mathcal{B}^o per e_o frame consecutivi, mentre la mano h rimane visibile, in quanto quando la mano esce dal campo visivo è probabile che stia ancora tenendo l'oggetto.

L'output del SOT produce un track τ_{o_i} spazio-temporale che segue la posizione dell'oggetto, ma non contiene informazioni sull'interazione stessa. A questo punto, o_i combina le informazioni relative alla durata temporale (start e end time) e ai bounding box spaziali dell'oggetto attivo, sfruttando sia la rilevazione HOI a livello di frame sia il tracciamento SOT.

Assignment and storing

Come definito in precedenza, ogni HOI tracklet o_i deve essere associato a una specifica istanza di oggetto. Per farlo, confrontiamo o_i con le istanze già presenti nella memoria.

In particolare, data l'insieme di HOI tracklets già memorizzati al tempo t , denotato \mathcal{O}_t , e l'insieme dei tracciamenti SOT osservati nello stesso istante, τ_t , verifichiamo se o_i possa essere associato a un'istanza esistente o se sia necessario creare una nuova. Per effettuare il confronto, calcoliamo prima le feature visive di o_i :

$$f(o_i) = \frac{1}{|\mathcal{V}_{o_i}|} \sum_{k \in \mathcal{V}_{o_i}} \gamma(k, b_k^o)$$

dove:

- \mathcal{V}_{o_i} è l'insieme dei frame associati a o_i ,
- b_k^o è il bounding box relativo al frame k ,
- γ visual-feature-extractor (nel nostro caso DINov2 [35]).

Per effettuare il *matching*, utilizziamo un approccio di clustering online basato su $f(o_i)$. La similarità tra o_i e una specifica istanza di oggetto id_j viene calcolata come:

$$s(o_i, id_j) = \frac{1}{|\mathcal{O}_t \in id_j|} \sum_{\mathcal{O}_t \in id_j} \langle f_{\mathcal{O}_t}, f_{o_i} \rangle$$

dove $\langle \cdot, \cdot \rangle$ indica la *cosine similarity* e $\mathcal{O}_t \in id_j$ è l'insieme dei tracklets associati all'istanza id_j .

Assegniamo o_i all'istanza id_j^* che massimizza la similarità e che supera una soglia θ (diversa dalla soglia utilizzata per l'IoU). Se un tracker in τ_t si

sovrappone significativamente con o_i e la sua confidenza è maggiore della similarità massima, allora o_i viene assegnato all'istanza del tracker. Altrimenti, viene assegnato a id_j^* . Qualora la similarità massima risultasse inferiore alla soglia, viene creata una nuova istanza per o_i .

Al termine di questa fase, le feature $f(o_i)$ e l'istanza assegnata vengono associate al tracklet o_i e memorizzate nella memoria \mathcal{E}

3.2.2 Location segments

Definiamo l'insieme dei *Location segments* \mathcal{L} come gli intervalli temporali durante i quali il soggetto svolge interazioni in *zone di attività principali*

Poiché un soggetto può interagire con più oggetti simultaneamente ma può trovarsi in un solo punto alla volta, ogni segmento $l_i \in \mathcal{L}$ viene modelizzato come un intervallo temporale che corrisponde all'inizio e alla fine di un'interazione in quella specifica zona.

Analogamente agli *HOI tracklets*, l'insieme \mathcal{L} viene costruito online.

Si seguendo due fasi principali: Temporal segmentation e Assignment and storing

Temporal segmentation

Dati i frame \mathcal{V}_t e le rilevazioni delle mani B_t^h , verifichiamo se la mano sta interagendo con un oggetto mentre si trova in una location. Per farlo, calcoliamo l'*optical flow*⁹ tra \mathcal{V}_{t-1} e \mathcal{V}_t e controlliamo la presenza di mani assicurandoci che $|B_t^h| > 0$.

Possiamo quindi determinare se il soggetto sta svolgendo un compito considerando le seguenti condizioni:

1. l'*optical flow* ha norma¹⁰ bassa;
2. è rilevata almeno una mano.

Questi criteri permettono di stabilire se il soggetto ha fatto una pausa (basso *optical flow*) ed è attivamente coinvolto nella scena (mano rilevata).

Analogamente agli HOI, applichiamo un filtraggio temporale, per cui un *Location segment* l_j è considerato attivo solo se entrambe le condizioni sono verificate per un numero consecutivo di frame s_l . Il segmento l_j viene terminato quando osserviamo un numero consecutivo di frame e_l in cui la norma dell'*optical flow* supera la soglia o non sono presenti mani rilevate.

⁹**Optical flow:** rappresenta il campo di movimento apparente dei pixel tra due frame consecutivi di un video, indicando la direzione e la velocità dello spostamento

¹⁰**Norma:** grandezza che rappresenta l'intensità complessiva di un vettore, nel nostro caso l'*optical flow*

Assignment and storing

Come per gli HOI, dobbiamo assegnare un'istanza alle location l_j definite temporaneamente. Agiamo analogamente, utilizzando però un *visual-feature-extractor* differente, σ (SWAG)[36].

Una volta ottenute le visual-feature g_{l_j} , calcoliamo la similarità tra tutte le istanze di location già presenti e assegniamo a l_j l'id che massimizza questa similarità, a condizione che superi una soglia prestabilita τ . Se la soglia non viene superata, viene creata una nuova istanza.

$$s(l_j, id_j) = \frac{1}{|\mathcal{L}_t \in id_j|} \sum_{\mathcal{L}_t \in id_j} \langle g_{\mathcal{L}_t}, g_{l_j} \rangle$$

Al termine di queste fasi, assegniamo $g(l_j)$ e l'istanza correlata nella nostra memoria \mathcal{E} .

3.3 Pseudocode

Di seguito vengono riportati i pseudocodici relativi alla costruzione della pipeline per la generazione degli elementi della memoria discussi nei paragrafi precedenti.

3.3.1 Object interactions

Algorithm 1 Object interactions pipeline

```

1: Input:
2:   Frames  $\{V_t\}$ 
3:   HOI detector  $\mathcal{D}$ 
4:   SOT tracker  $\mathcal{J}$ 
5:   Similarity threshold  $\theta$ 
6: Output:
7:   Set of hand-object interaction tracklets  $\mathcal{O}$ 
8: for each frame  $V_t$  do do
9:    $\mathcal{B}_t^o, \mathcal{B}_t^h \leftarrow \mathcal{D}(V_t)$  (Detect hands and objects)
10:  for each detection  $(b^o, b^h) \in (\mathcal{B}_t^o, \mathcal{B}_t^h)$  do do
11:    if new hand-object interaction (i.e. no detections in the last  $w_s$  frames) then then
12:      Create new tracklet  $o_i$ 
13:      Start SOT  $\mathcal{J}_{o_i}$  for  $o_i$ 
14:    end if
15:  end for
16:  for each tracklet  $o_i$  do do
17:    Update the detections with  $\mathcal{J}_{o_i}$ 
18:    if  $\exists \tilde{b}^h \in \mathcal{B}_t^h$  matching with  $o_i$  in the last  $e_o$  frames and  $|\tilde{b}^h| > 0$  then then
19:      Mark  $o_i$  as complete
20:    end if
21:  end for
22: end for
23: for each completed tracklet  $o_i$  do do
24:   Compute visual features  $f(o_i)$  (Eqn. 1)
25:   Compute similarity  $s(o_i, id_j)$  with existing instances in  $\mathcal{O}$  (Eqn. 2)
26:   if maximum similarity  $> \theta$  then then
27:     Assign  $o_i$  to best matching instance  $id_j$ 
28:   else
29:     Create new instance for  $o_i$ 
30:   end if
31:   Store  $o_i$  in  $\mathcal{O}$ 
32: end for
33: return  $\mathcal{O}$ 
  
```

3.3.2 Location Segment

Algorithm 2 Location Segment pipeline

```

1: Input:
2:   Frames  $\{V_t\}$ 
3:   HOI detector  $\mathcal{D}$ 
4:   Similarity threshold  $\tau$ 
5: Output:
6:   Set of location segments  $\mathcal{L}$ 
7: for each frame  $V_t$  do
8:    $\mathcal{B}_t^o, \mathcal{B}_t^h \leftarrow \mathcal{D}(V_t)$  (Detect hands and objects)
9:   Compute optical flow OpticalFlow( $V_{t-1}, V_t$ )
10:  if location segment  $l_j$  is active then then
11:    if high |OpticalFlow( $V_{t-1}, V_t$ )| or  $|\mathcal{B}_t^h| = 0$  for  $e_l$  consecutive frames
     then then
12:      Mark  $l_j$  as complete
13:    else
14:      Continue  $l_j$ 
15:    end if
16:  else
17:    if low |OpticalFlow( $V_{t-1}, V_t$ )| and  $|\mathcal{B}_t^h| > 0$  for  $s_l$  consecutive
     frames then then
18:      Subject is interacting, start active location segment  $l_j$ 
19:    end if
20:  end if
21: end for
22: for each completed segment  $l_j$  do
23:   Compute visual features  $g(l_j)$ 
24:   Compute similarity  $s(l_j, id_i)$  with existing instances in  $\mathcal{L}$ 
25:   if maximum similarity  $> \tau$  then then
26:     Assign  $l_j$  to best matching instance  $id_i$ 
27:   else
28:     Create new instance for  $l_j$ 
29:   end if
30:   Store  $l_j$  in  $\mathcal{L}$ 
31: end for
32: return  $\mathcal{L}$ 
  
```

3.4 AMB - Active Memories Benchmark

Per studiare l'interazione tra oggetti attivi, location e la loro correlazione. è stato introdotto un benchmark ad-hoc, l'*Active Memories Benchmark* (AMB).

Il benchmark comprende 20.500 query che coprono diversi livelli di ragionamento. Le query sono formulate come domande a scelta multipla.

Ad esempio, alcune domande riguardano l'utilizzo di oggetti: “Quale oggetto ho usato con [VQ]?” dove [VQ] rappresenta un ritaglio visivo dell’oggetto; altre domande chiedono “In quali location ho usato [VQ]?”

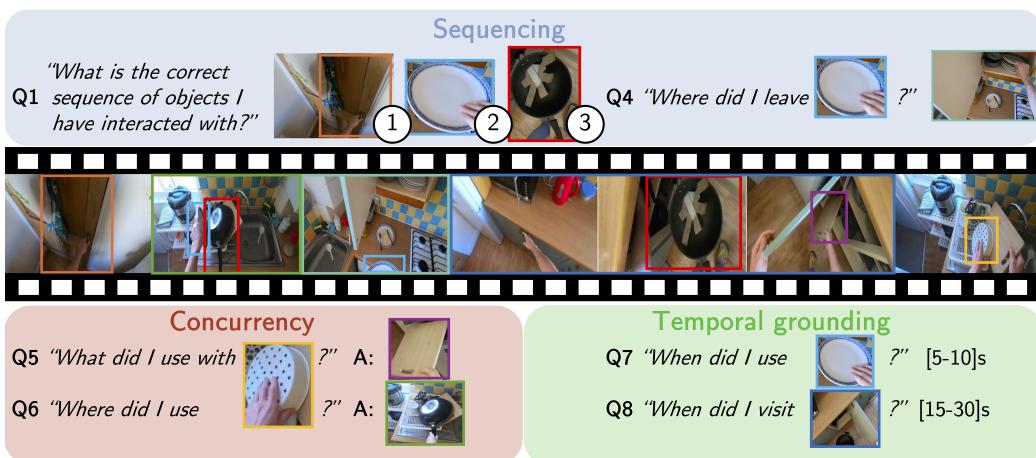


Figura 3.3: Esempi di query del benchmark AMB.

Come si nota, nelle domande non compaiono i nomi degli oggetti. Ogni visual query di un oggetto (VQ), risposta visiva dell’oggetto (VA), query sulla location (LQ) o risposta sulla location (LA) è parametrizzata tramite patch visive.

3.4.1 Tipologie di query

Le query sono strutturate in tre macro aree:

- **Sequencing (SQ) [Q1-4]:** valutano la capacità di discriminare l’ordine temporale degli eventi. Ad esempio, il modello deve ordinare le interazioni nel tempo e identificare quale oggetto è stato utilizzato prima o dopo un altro.
- **Concurrency (CO) [Q5-6]:** valutano la capacità di catturare interazioni multiple simultanee. Ad esempio, verificare se diversi oggetti sono stati utilizzati insieme (oggetto-oggetto) o se un oggetto è stato usato in una specifica location (oggetto-location).

- **Temporal Grounding (TG) [Q7-8]:** valutano la capacità di recuperare tutti gli intervalli temporali in cui un oggetto o una location è stato coinvolto in interazioni.

Reasoning	Query	Template	Dim.	Answer	Qs
SQ	Q1	What is the correct sequence of objects I have interacted with?	O	Obj. seqs	464
	Q2	What did I use with the left/right hand after [VQ]?	O	Obj.	3466
	Q3	What did I use with the left/right hand before [VQ]?	O	Obj.	3466
	Q4	Where did I take/leave [VQ]?	O, L	Loc.	1266
CO	Q5	What did I use with [VQ]?	O	Obj. sets	2105
	Q6	Where did I use [VQ]?	O, L	Loc. sets	2320
TG	Q7	When did I use [VQ]?	O	Intervals	2614
	Q8	When did I visit [LQ]?	O, L	Intervals	809

Tabella 3.1: Tipologia di domande AMB con colonne colorate per Reasoning e Query.

3.4.2 Risposte alle Query

approfondire come si elabora la risposta alle query

3.5 Risultati

AMEGO è stato valutato confrontandolo con diverse baseline comuni per il task di video-QA:

- **Semantic-free QA (SF-QA)** utilizza modelli vision-language, come CLIP [37], per mappare query, video e risposte nello stesso spazio di embedding. Le feature visive vengono estratte dai frame del video, dai patch delle query e dalle risposte, mentre le feature testuali provengono dalla domanda. L'embedding della query è ottenuto come media delle feature, e la risposta con la similarità più alta viene selezionata.
- **SF-QA (obj)** è una variante di SF-QA che include anche le feature visive degli oggetti attivi rilevati da [33].
- **Semantic QA (S-QA)** sfrutta "captioner" pre-addestrati per generare un sommario semantico del video. Si usano LaViLa [38] per il video egocentrico e BLIP-2 [39] per i patch della query. Le caption vengono poi passate a LLaMA-2-7B [40] per rispondere alle domande. Se il testo supera i 4096 token, viene sottocampionato.

- **Multi-round semantic QA (LLoVi)** [41] funziona in due round: prima sintetizza le caption del video alla luce della domanda, poi risponde alla query usando il sommario generato.

I risultati sono riportati per AMEGO si dividono in: AMEGO-S e AMEGO-L, a seconda della dimensione del visual feature extractor (ViT-S/B vs ViT-L).

Tabella 3.2: Accuracy (%) sulle diverse query di AMB. Migliori valori in grassetto.

Method	SQ				CO		TG		Total
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	
Random	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
SF-QA	13.7	21.6	22.5	26.8	22.1	31.9	23.7	26.2	22.0
SF-QA (obj)	13.1	23.4	22.6	23.2	21.7	26.1	23.8	25.2	21.2
S-QA (LaViLa)	20.9	20.6	21.2	24.6	24.9	27.1	21.4	22.6	22.4
S-QA (BLIP-2)	23.9	22.0	22.5	23.3	27.5	27.0	20.2	24.1	23.6
S-QA (LaViLa+BLIP-2)	22.8	22.2	21.4	22.6	25.1	26.1	21.4	24.5	22.9
LLoVi (LaViLa)	21.1	20.2	20.8	21.0	21.2	20.3	20.5	21.6	20.8
LLoVi (BLIP-2)	22.3	21.4	21.8	22.2	25.6	26.7	18.1	22.2	22.4
LLoVi (LaViLa+BLIP-2)	22.8	21.9	21.5	24.6	25.3	26.5	18.5	19.8	22.6
AMEGO - S	32.0	35.1	34.8	35.8	24.7	37.8	33.6	44.3	33.8
AMEGO - L	33.7	36.3	37.2	38.3	27.6	44.3	34.7	48.9	36.3

Tutte le baseline ottengono risultati migliori sulle query relative alla *currency*, probabilmente perché sfruttano pattern ricorrenti presenti nei dati di addestramento, ad esempio una padella spesso utilizzata sul piano cottura [2]. Tuttavia, le performance complessive rimangono vicine alla soglia della scelta casuale.

AMEGO invece ottiene buoni risultati su tutte le tipologie di query, superando le baseline con un margine consistente (+12.7%). La domanda in cui AMEGO mostra maggior difficoltà è Q5, a causa dei limiti attuali dei detector di interazione mano-oggetto nel predire oggetti multipli che interagiscono contemporaneamente con la stessa mano del soggetto.

Capitolo 4

Dataset

- epic - whats - come è stato acquisito - cosa viene dato e in che formato -
- stats - come è stato usato durante la creazione di AMB
- enigma //

Capitolo 5

Esperimenti

- spiegare il setup degli esperimenti (pag. 10-11)
 - replica risultati epic kitchen - descrizione macchina old - dettagli tecnici - costruzione dockerfiles - amego - epic kitchen download - build dockerfile su vecchia macchina - i problema scheda video troppo vecchia - passaggio alla nuova - dettagli tecnici - build docker files - run scripts su epic (optical flow, hand object detection, interaction tracklets, location segments) - descrizione cosa fanno questi file
 - passaggio ad enigma - scaricamento files - frame - annotazioni - video (usati per estrarre i frame da usare negli script amego) (i frame scaricabili dal sito non erano tutti ma solo quelli annotati nel json)
 - run scripts
 - creazione benchmark (spiegare come sono stati costruiti i json) - q5
 - q6

Capitolo 6

Risultati

Conclusione

Bibliografia

- [1] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. Egocentric vision-based action recognition: A survey. *Neurocomputing*, 472:175–197, 2022.
- [2] Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. Amego: Active memory from long egocentric videos. In *European Conference on Computer Vision*, 2024.
- [3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021.
- [4] Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, and Giovanni Maria Farinella. Enigma-51: Towards a fine-grained understanding of human-object interactions in industrial scenarios. *IEEE Winter Conference on Application of Computer Vision (WACV)*, 2024.
- [5] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023.
- [6] Xitong Yang, Fu-Jen Chu, Matt Feiszli, Raghav Goyal, Lorenzo Torresani, and Du Tran. Relational space-time query in long-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6398–6408, June 2023.
- [7] Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding, 2024.

- [8] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa, 2025.
- [9] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent, 2024.
- [10] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos, 2024.
- [11] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos, 2025.
- [12] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition, 2022.
- [13] Yunxin Li, Xinyu Chen, Baotain Hu, and Min Zhang. Llms meet long video: Advancing long video question answering with an interactive visual adapter in llms, 2024.
- [14] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models, 2024.
- [15] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding, 2024.
- [16] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2024.
- [17] Anurag Arnab, Chen Sun, and Cordelia Schmid. Unified graph structured models for video understanding, 2021.
- [18] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos, 2018.

- [19] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation, 2021.
- [20] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs, 2016.
- [21] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as composition of spatio-temporal scene graphs, 2019.
- [22] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding, 2018.
- [23] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network, 2018.
- [24] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs, 2018.
- [25] Will Price, Carl Vondrick, and Dima Damen. Unweavenet: Unweaving activity stories, 2022.
- [26] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos, 2023.
- [27] Yong Jae Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. pages 1346–1353, 06 2012.
- [28] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.
- [29] Bo Xiong and Kristen Grauman. Detecting snap points in egocentric video with a web photo prior. pages 282–298, 09 2014.
- [30] Yen-Liang Lin, Vlad I. Morariu, and Winston Hsu. Summarizing while recording: Context-based highlight detection for egocentric videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

- [31] Bin Zhao and Eric P. Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [32] Bo Xiong, Gunhee Kim, and Leonid Sigal. Storyline representation of egocentric videos with an applications to story-based search. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [33] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale, 2020.
- [34] Hao Tang, Kevin Liang, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset, 2023.
- [35] Maxime Oquab, Timothée Darzet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.
- [36] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models, 2022.
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [38] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models, 2022.
- [39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
- [40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal

Bhargava, Shruti Bhosale, Dan Bikell, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

- [41] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering, 2024.