

From ScanDDM to ART: DDM-DINO

Hari Calzi - 52061A

Abstract—I modelli di attenzione visuale hanno mostrato una crescente capacità nel predire gli scanpath, ovvero le sequenze di fissazioni e movimenti oculari. In particolare, ScanDDM [1] ha introdotto un approccio basato su DDM per la predizione di *scanpath goal-directed* in modalità *zero-shot*, mentre ART [2] si è focalizzato sulla predizione incrementale dell'attenzione durante compiti di *object referral* guidati dal linguaggio. Il presente lavoro esplora la combinazione di questi due approcci, modificando ScanDDM con l'integrazione di GroundingDINO [3] per affrontare il compito di *incremental object referral*. Il modello risultante è stato denominato DDM-DINO.

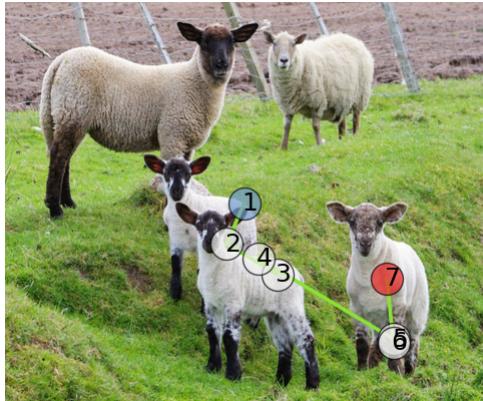


Fig. 1: "Small sheep on right front"

1 INTRODUZIONE

Comprendere i meccanismi alla base dell'attenzione visuale umana è una sfida fondamentale in molteplici discipline, dalle neuroscienze cognitive all'intelligenza artificiale. La capacità di analizzare e predire come gli esseri umani dirigono il proprio sguardo è cruciale per una vasta gamma di applicazioni, relative all'interazione uomo-macchina. Tradizionalmente, i modelli computazionali di attenzione visuale si sono concentrati principalmente su proprietà fisiche delle immagini, come la salienza. Tuttavia, è sempre più evidente che l'attenzione è fortemente influenzata da fattori cognitivi di alto livello, come gli obiettivi e il linguaggio. Questo lavoro si inserisce in questa direzione, esplorando un nuovo approccio per modellare l'attenzione guidata dal linguaggio.

H. Calzi, Corso di Natural Interaction & Affective Computing, A/A 2024-2025, Università degli Studi di Milano, via Celoria 18, Milano, Italy
E-mail: hari.calzi@studenti.unimi.it

1.1 Importanza del problema

L'interazione naturale uomo-macchina richiede sistemi in grado di comunicare con gli utenti in modo intuitivo ed efficace. La predizione dello scanpath umano gioca un ruolo chiave in questo, permettendo ai sistemi di anticipare dove l'utente sta guardando e, di conseguenza, di fornire informazioni o assistenza in modo più pertinente. Ad esempio, nei sistemi di guida assistita la predizione dell'attenzione può migliorare la sicurezza, consentendo al sistema di fornire avvisi tempestivi e di anticipare le intenzioni del conducente. Inoltre, la modellizzazione dell'attenzione guidata dal linguaggio è essenziale per sviluppare assistenti virtuali più sofisticati, capaci di comprendere e rispondere alle istruzioni degli utenti in modo naturale e contestuale.

1.2 Approccio seguito

Inizialmente è stata condotta un'analisi approfondita di ScanDDM e ART, sia a livello teorico che implementativo. Da tale analisi sono state estratte le principali divergenze tra le architetture.

Nello specifico, a livello di funzionamento, entrambi i modelli accettano in input una rappresentazione visiva (immagine) e un'espressione linguistica e forniscono in output una predizione dello scanpath umano sovrapposto all'immagine. ART, in particolare, è progettato per affrontare il compito di *incremental object referral*, ovvero la predizione progressiva dell'attenzione visiva di un osservatore mentre ascolta una descrizione linguistica che si riferisce a un oggetto specifico all'interno della scena.

A livello teorico invece si evidenziano le seguenti distinzioni: ART impiega un meccanismo di *contextual embedding* della sequenza linguistica di input, integrando l'informazione derivante dalle parole precedenti per modulare la rappresentazione della parola corrente, mentre ScanDDM processa l'intera espressione linguistica come un'entità atomica. Inoltre, ScanDDM opera in modalità *zero-shot*, generalizzando a nuove descrizioni linguistiche senza richiedere un addestramento specifico, a differenza di ART che è stato addestrato per predire l'attenzione incrementale. Sebbene entrambi i modelli condividano componenti architettoniche comuni, quali moduli di *feature extraction* visuale e linguistica, ART si avvale di un *autoregressive transformer decoder* per

la predizione incrementale delle fissazioni, elemento non presente in ScanDDM.

Al fine di colmare il divario, ScanDDM è stato modificato con l'integrazione di GroundingDINO [3], un modello di *object grounding*, per emulare il comportamento incrementale di ART, preservando la sua natura zero-shot e una struttura architetturale più parsimoniosa.

1.3 Contributi

A seguito della fase di testing di entrambi i modelli, ART e ScanDDM, si è intrapreso lo sviluppo di DDM-DINO. La prima modifica sostanziale ha riguardato l'integrazione di GroundingDINO [3], un modello avanzato di *object grounding*, nello specifico la variante *IDEA-Research/grounding-dino-base*. In contrasto, ScanDDM impiegava un modello di *clip segmentation* denominato *CIDAS/clipseg-rd64-refined* per la segmentazione delle immagini. Diversamente dall'*object grounding*, che ha dimostrato elevata efficacia nella localizzazione di oggetti target mediante descrizioni verbali dettagliate, la *clip segmentation* tendeva a generare segmentazioni meno specifiche e con una ridotta sensibilità alla posizione degli oggetti, in particolare in presenza di istanze multiple o qualora le descrizioni contenessero riferimenti spaziali.

Un'illustrazione di quanto esposto è osservabile nella Fig. 2, che raffigura un gregge di pecore. La *bounding box* in rosso evidenzia l'oggetto di ricerca, ottenuto in questo caso fornendo il prompt "*small sheep on right front*". Nella Fig. 3 si può notare come, utilizzando ClipSeg, l'area di maggiore interesse appaia concentrata nella porzione superiore dell'immagine, intersecando la recinzione e le pecore. Al contrario, utilizzando GroundingDINO, come mostrato nella Fig. 4, la mappa di salienza corrisponde precisamente all'oggetto atteso, sovrapponendosi al riquadro di delimitazione presente nella Fig. 2.

Tale implementazione ha consentito l'ottenimento di mappe di salienza più precise, pur presentando alcune

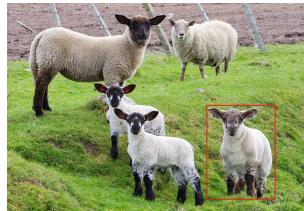


Fig. 2: "Small sheep on right front"

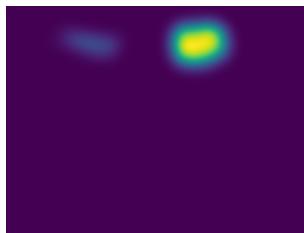


Fig. 3: "Salienza con ClipSeg"

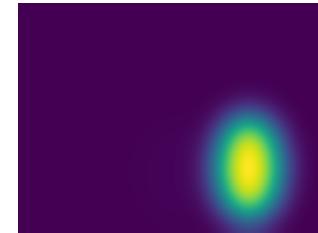


Fig. 4: "Salienza con GroundingDINO"

sfide nella loro generazione. La problematica affrontata risiede nella differente modalità di output di GroundingDINO, che, a differenza di CLIPSeg, restituisce le coordinate del bounding box pertinente all'oggetto identificato tramite la descrizione fornita. Originariamente, in ScanDDM, il modello CLIPSeg generava logit per ciascun pixel, convertiti successivamente in mappe di probabilità mediante l'applicazione di una funzione sigmoide, quantificando la corrispondenza con l'input testuale. Queste mappe di probabilità venivano quindi aggregate per costituire una mappa di salienza iniziale. Al fine di simulare la tendenza a formare mappe di salienza composte da valori sfumati nell'intervallo [0,1], è stata implementata la creazione di un'ellisse inscritta all'interno del bounding box. La maschera ellittica viene sommata alla mappa di salienza preesistente, emulando un'intensificazione dell'attenzione all'interno della regione relativa al soggetto ricercato. La mappa di salienza risultante è infine sottoposta a un processo di smoothing mediante un filtro gaussiano e successivamente normalizzata. Un esempio è riportato in Fig. 4.

Al fine di simulare il processo di incremental object referral e l'embedding contestuale di ART, si è adottata una strategia iterativa su ScanDDM. In ciascuna iterazione al modello viene fornita la medesima immagine unitamente a una mappa di salienza composita, risultante dalla ponderazione delle mappe precedenti e di quella relativa al prompt corrente. Il prompt viene fornito in modalità incrementale (es. "small", "small sheep", "small sheep on", ecc.), consentendo al modello di basare la sua analisi sull'input visivo corrente, integrando al contempo le informazioni elaborate nelle iterazioni precedenti. Per emulare il funzionamento di ART, il prompt è stato integrato con i token speciali *begin of talk* (BOT) ed *end of talk* (EOT). Questi rappresentano due intervalli temporali distinti: il primo, BOT, simula la fase iniziale in cui l'osservatore esamina l'immagine in assenza di riferimenti testuali. Il secondo, EOT, analogamente, simula la fase finale successiva alla completa enunciazione dei riferimenti testuali.

Lo scanpath generato da DDM-DINO è costituito dalla fissazione dominante di ogni iterazione del DDM, permettendo di analizzare la regione dell'immagine su cui l'osservatore concentra la propria attenzione in relazione al termine udito, analogamente ad ART.

Le prestazioni del modello risultante sono state valutate sul dataset RefCOCO-Gaze [2] mediante un insieme

di metriche descritte nella sezione 5.

2 ANALISI DELLO STATO DELL'ARTE

La modellizzazione dell'attenzione visiva ha visto progressi significativi negli ultimi anni, passando dai primi modelli focalizzati sulla salienza [4] ad approcci più complessi che predicono gli scanpath [5]. Mentre la ricerca iniziale si è spesso concentrata sulle condizioni di visione libera [6], c'è un crescente interesse per l'attenzione guidata da obiettivi, in cui l'allocazione dell'attenzione è influenzata da compiti specifici come la ricerca visiva e la didascalia di immagini. In questo contesto, questo progetto si basa principalmente su due studi chiave.

In primo luogo, ScanDDM [1] introduce una nuova prospettiva inquadrandola l'attenzione guidata da obiettivi come un processo decisionale percettivo, utilizzando un *multialternative Drift Diffusion Model* (DDM) in un contesto zero-shot. Questo approccio si distingue per la sua capacità di modellare l'allocazione dell'attenzione come un processo decisionale continuo, in linea con le teorie neurobiologiche.

In secondo luogo, il modello Attention in Referral Transformer (ART) [2] si distingue per aver affrontato specificamente il compito dell'incremental object referral, un'area meno esplorata nella ricerca sull'attenzione computazionale. ART affronta le sfide uniche di questo compito, come l'integrazione di informazioni linguistiche e visive incremental.

Questi ultimi due modelli forniscono le basi teoriche e metodologiche per questo studio.

3 MODELLO TEORICO

In quanto DDM-DINO deriva la sua architettura da ScanDDM, integrandola con GroundingDINO, la presente sezione è dedicata all'analisi dei principali modelli teorico-matematici impiegati in ScanDDM e GroundingDINO, presi dai rispettivi paper.

3.1 ScanDDM

Si consideri l'immagine come un insieme discreto di *patch*, identificate con i pixel p che la compongono, rappresentando le possibili locazioni su cui l'osservatore può dirigere l'attenzione. A ciascuna patch p è assegnato un valore a priori V_p , determinato in funzione del compito visivo. Tale valore è definito dalla seguente relazione:

$$V_p = \rho(p \in \mathbb{O}_G) \quad (1)$$

L'Eq. 1 esprime il valore V_p associato a ogni patch p come la probabilità che p appartenga a un oggetto di interesse \mathbb{O}_G , condizionata dall'obiettivo corrente G dell'osservatore. In ScanDDM, questa probabilità è determinata a partire dalla mappa di segmentazione dell'oggetto di interesse. Nel presente lavoro, tale stima è stata raffinata attraverso un processo di object grounding, come illustrato nella sezione 1.3.

Il processo decisionale a livello neurobiologico è modellato da popolazioni neuronali che accumulano evidenze con rumore, finché una soglia di attivazione non viene superata. In scenari semplificati con due sole opzioni e un'evoluzione continua nel tempo, il DDM fornisce una formalizzazione di questo processo. Il vanilla DDM inizia con evidenza pari a zero e la accumula seguendo una *Stochastic Differential Equation* (SDE) della forma:

$$dq(t) = Idt + \sigma dW(t) \quad (2)$$

dove $dq(t)$ rappresenta l'evidenza accumulata nell'intervallo di tempo dt , Idt viene detto *constant drift term* e rappresenta l'incremento medio di evidenza a supporto di una scelta per unità di tempo, $\sigma dW(t)$ invece rappresenta il rumore come una distribuzione Gaussiana con media 0 e varianza σdt . L'evidenza quindi aumenta nel tempo con una velocità media I , ma è anche influenzata dall'accumulazione del rumore. La decisione viene presa quando q raggiunge la soglia, a o $-a$, e il tempo di reazione è determinato dal tempo necessario per raggiungere tale soglia, ovvero la durata di una fissazione.

Quando si considerano molteplici alternative decisionali si impiega un *race model*: si formula un SDE per ciascuna alternativa e il processo decisionale è modellato come una competizione tra le alternative per il raggiungimento della soglia di decisione a . Ogni alternativa integra l'evidenza in modo indipendente:

$$dq_p(t) = I_p dt + \sigma dW(t), \quad p = 1, \dots, N_P. \quad (3)$$

La decisione finale nel processo decisionale umano è considerata co-dipendente. In ogni istante, il cervello calcola un *Relative Decision Value* (RDV) per ciascuna delle possibili alternative. I RDV sono vincolati in modo tale che la loro somma sia nulla in ogni istante temporale:

$$RDV_p(t) = q_p(t) - \max[q_k(t)], \quad \forall k \neq p \quad (4)$$

I_p , definito come *drift rate*, è associato alla competizione relativa alla patch p -esima (vedi Eq. 3) e dipende da V_p , il valore a priori, e da V_{p^*} , il valore della patch attualmente esplorata. I_p è definito come:

$$I_p = \eta \cdot \frac{\psi(p, p^*)}{1 + \log_2(1 + V_{p^*})} \quad (5)$$

Nell'Eq 5, η rappresenta una costante positiva che definisce la velocità di accumulo di evidenza per ciascuna patch. Il denominatore scala il drift rate di tutte le patch in funzione del valore a priori della patch corrente. La funzione ψ , definita come *gazing function*, è espressa come segue:

$$\psi(p, p^*) = V_p \cdot A(p, p^*) \cdot D(p, p^*) \quad (6)$$

La gazing function in Eq. 6 è definita come il prodotto di tre termini. V_p rappresenta il valore a priori di ciascuna patch, mentre $A(p, p^*)$ è definito come:

$$A(p, p^*) = \frac{1}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2} \right] \quad (7)$$

$A(p, p^*)$ impiega una distribuzione di Cauchy per ponderare la visibilità della patch p rispetto alla patch corrente p^* . Questa scelta metodologica privilegia l'esplorazione delle regioni adiacenti, pur mantenendo una sensibilità verso i pixel più distanti, una caratteristica intrinseca della distribuzione di Cauchy grazie alla sua forma a code lunghe. $D(p, p^*)$, invece, è definita come segue:

$$D(p, p^*) = \frac{\cos(\alpha \cdot \theta(p, p^*)) + 1}{2} \quad (8)$$

L'Eq. 8 incorpora la tendenza a favorire l'inizio di saccadi nelle direzioni orizzontali e verticali rispetto a quelle oblique, dove $\theta(p, p^*)$ quantifica l'angolo, espresso in gradi, tra le posizioni p e p^* .

3.2 GroundingDINO

GroundingDINO processa coppie (Immagine, Testo) fornendo le bounding box e le corrispondenti espressioni testuali dei target della ricerca. L'architettura, basata su due *encoder* e un *decoder*, estrae separatamente caratteristiche da immagini e testi, per poi fonderle. Un modulo guidato dal linguaggio seleziona aree di interesse nell'immagine, che vengono ulteriormente elaborate dal decoder per raffinare le localizzazioni degli oggetti e le loro descrizioni testuali. Il sistema analizza l'immagine, a diverse risoluzioni e il testo, utilizzando meccanismi di attenzione per allineare le informazioni visive e linguistiche.

Per sfruttare efficacemente l'informazione testuale in ingresso al fine di guidare il processo di rilevamento degli oggetti, si introduce un modulo di *Language-Guided Query Selection*. Questo modulo ha lo scopo di identificare e selezionare, dalle *feature* dell'immagine codificata, quelle che presentano una maggiore rilevanza rispetto al testo fornito come input, destinandole a fungere da query per il decoder. Siano $\mathbf{X}_I \in R^{N_I \times d}$ le feature dell'immagine e $\mathbf{X}_T \in R^{N_T \times d}$ le feature del testo, dove N_I rappresenta il numero di token dell'immagine, N_T il numero di token del testo, d la dimensione dello spazio delle feature. L'obiettivo è estrarre le N_q query più rilevanti dalle feature dell'immagine codificata per alimentare il decoder. Gli indici \mathbf{I}_{N_q} delle N_q query più rilevanti vengono determinati attraverso la seguente espressione:

$$\mathbf{I}_{N_q} = \text{Top}_{N_q}(\text{Max}^{(-1)}(\mathbf{X}_I \mathbf{X}_T^\top)) \quad (9)$$

Nell'Eq. 9 l'operatore Top_{N_q} seleziona i primi N_q indici in base al valore, la funzione $\text{Max}^{(-1)}$ invece esegue l'operazione di massimo lungo la dimensione -1 , mentre \top indica la trasposizione matriciale. A partire dagli N_q indici selezionati si estraggono le feature corrispondenti per l'inizializzazione delle query del decoder. L'inizializzazione delle query del decoder si avvale di una strategia di selezione mista, in cui ogni query è costituita da due componenti separate: una relativa al contenuto e una di natura posizionale. La componente

di contenuto è definita come un insieme di parametri che vengono appresi durante la fase di addestramento del modello. La componente posizionale invece è implementata tramite un insieme di *anchor box* dinamici, i cui valori iniziali sono derivati dalle uscite dell'encoder.

4 SIMULAZIONE E ESPERIMENTI

Il modello proposto, DDM-DINO, è stato sottoposto a una fase di valutazione sperimentale mediante l'utilizzo di svariate coppie immagine-testo. Tale sperimentazione è stata condotta al fine di accertare l'efficacia del modello nel riprodurre il comportamento dello scanpath umano in risposta a stimoli visuo-linguistici complessi.

4.1 Dataset

Il dataset utilizzato è RefCOCO-Gaze [2]. Il dataset comprende 19.738 scanpath registrati durante la partecipazione di 220 individui con visione normale o corretta. I partecipanti hanno osservato 2.094 immagini tratte dal dataset COCO, ascoltando simultaneamente le espressioni referenziali associate provenienti dal dataset RefCOCO. I dati di tracciamento oculare, acquisiti tramite un *eyetracker EyeLink 1000*, includono informazioni dettagliate sulla posizione e la durata di ciascuna fissazione, il bounding box dell'oggetto target della ricerca, le registrazioni audio delle espressioni referenziali, la precisa tempistica della parola target all'interno dell'espressione e la sincronizzazione temporale tra le parole pronunciate e la sequenza di fissazioni oculari (indicando quale parola ha innescato specifiche fissazioni).

4.2 Architettura del sistema

Il diagramma illustrato in Fig. 5 presenta lo schema concettuale dell'architettura di DDM-DINO, derivato dall'analogo schema proposto nel lavoro di ScanDDM [1]. Il modello acquisisce in ingresso lo stimolo visivo (immagine) e la descrizione testuale (prompt). Il prompt viene segmentato a livello di parola e processato in maniera incrementale, come descritto dettagliatamente nella sezione 1.3. DDM-DINO esegue un processo iterativo su ciascuna porzione incrementale del prompt, applicando il modello GroundingDINO per generare, a partire dall'immagine e dal segmento di prompt corrente, una mappa di salienza.

Il funzionamento interno di GroundingDINO è riportato per completezza, sebbene non abbia rappresentato il focus primario dell'analisi condotta nel presente progetto. Architetturalmente, GroundingDINO include due *backbone* distinte, deputate all'estrazione delle feature dell'immagine e del testo. Tali feature sono successivamente elaborate e raffinate attraverso un modulo di *feature enhancement*. Il risultato di questo processo è inoltrato al modulo di language-guided query selection, illustrato nella sezione 3.2. Il decoder produce infine la bounding box, espressa in coordinate spaziali, dell'oggetto di interesse all'interno dell'immagine. A

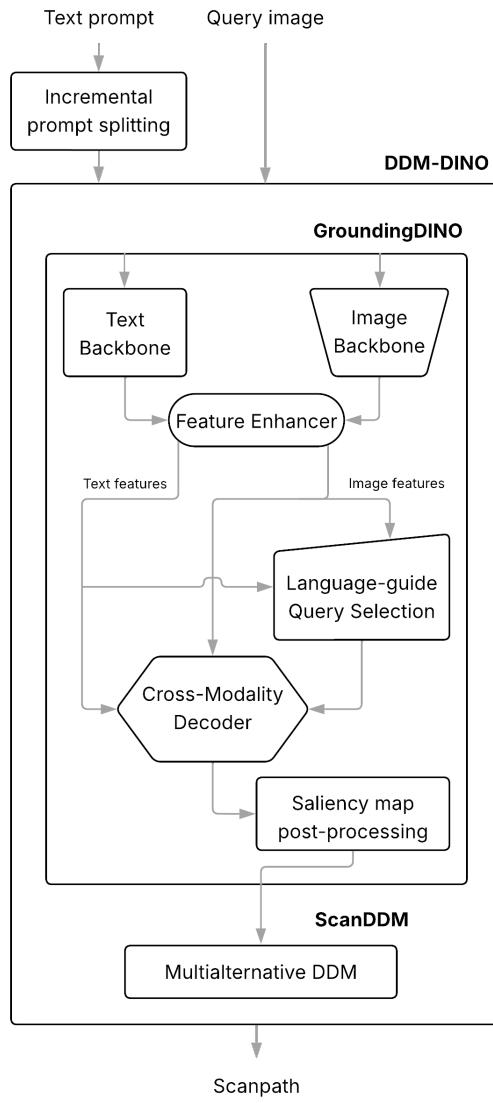


Fig. 5: Architettura funzionale del sistema

partire da tale bounding box viene generata la mappa di salienza, secondo la procedura descritta nella sezione 1.3. Questa mappa di salienza costituisce l'input del modello DDM originale di ScanDDM, il quale simula il comportamento umano e genera lo scanpath. Come precedentemente illustrato, tale procedimento viene iterato sequenzialmente fino al completo processamento del prompt. DDM-DINO fornisce in output lo scanpath finale risultante dall'intero processo. Alcuni esempi di output generati dal modello sono presentati nella sezione 5.1.

4.3 Dettagli implementativi

I tre modelli oggetto di analisi, ScanDDM, DDM-DINO e ART, sono stati valutati impiegando le 92 immagini e le relative descrizioni testuali contenute nello split di validazione del dataset precedentemente descritto

nella sezione 4.1. A partire dagli scanpath generati da ciascuna esecuzione dei tre modelli è stato calcolato il valore delle metriche selezionate, illustrate nella sezione 5.2, utilizzando come riferimento comparativo i dati di tracciamento oculare presenti nello stesso split di validazione del dataset umano.

5 RISULTATI OTTENUTI

5.1 Esempi di funzionamento

Nella presente sezione sono illustrati tre esempi operativi di DDM-DINO, confrontando gli scanpath generati da DDM-DINO con quelli prodotti da ScanDDM e ART e con quelli registrati dall'essere umano. Le immagini e i relativi prompt impiegati per tale comparazione sono stati selezionati dal dataset descritto nella sezione 4.1 e dal lavoro di ART [2].

5.1.1 Esempio 1: pecore

Il primo caso di studio analizzato è relativo al prompt *"small sheep on right front"*. Lo stimolo visivo ritrae un gregge di pecore in un prato, con alcuni esemplari più giovani in primo piano.

La Fig. 7 illustra lo scanpath di riferimento prodotto dal modello ART, configurandosi come l'obiettivo prestazionale, il quale presenta una discreta somiglianza con lo scanpath umano osservabile in Figura 6. Entrambi gli scanpath si concentrano inizialmente sulla pecora al centro dell'immagine e poi si spostano dolcemente verso la pecora a destra.

DDM-DINO, come evidenziato nella Figura 8, concentra inizialmente l'attenzione prevalentemente sulla pecora centrale, che rientra comunque nella descrizione iniziale di *"small sheep"*. Successivamente, con l'aggiunta dell'espressione *"on right front"* al prompt, il modello adatta la sua attenzione, convergendo sulla pecora situata sulla destra.

In questo scenario il modello di base ScanDDM, come mostrato nella Figura 9, genera uno scanpath che si discosta significativamente dagli altri presentati, focalizzando la sua attenzione su regioni differenti dell'immagine. Tale comportamento è riconducibile alla generazione di una mappa di salienza non accurata, come descritto nella sezione 1.3 e illustrato nella Fig. 3.

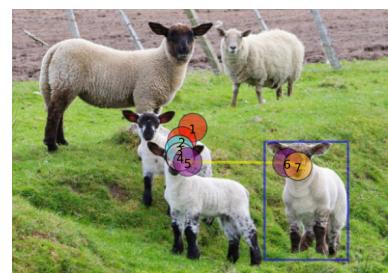


Fig. 6: "Pecore, umano"

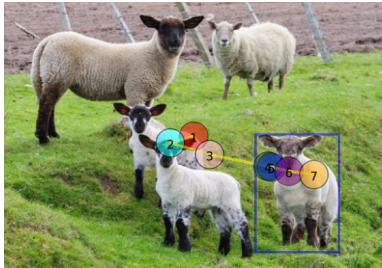


Fig. 7: "Pecore, ART"

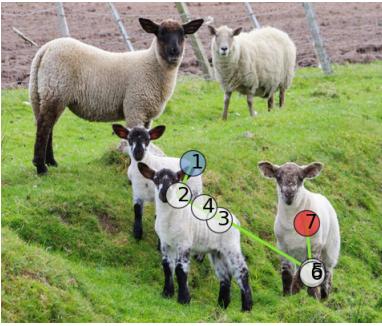


Fig. 8: "Pecore, DDM-DINO"

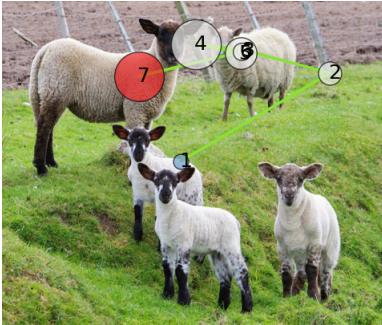


Fig. 9: "Pecore, ScanDDM"

5.1.2 Esempio 2: persone

Nel secondo caso di studio, il prompt testuale analizzato è *"standing girl in pink left of photo"*. Lo stimolo visivo ritrae un adulto e alcuni bambini equipaggiati con abbigliamento invernale, presumibilmente durante una lezione di sci.

In Fig. 10 viene illustrato lo scanpath umano registrato durante le sessioni di test. Questo dato è rilevante in quanto evidenzia come anche il modello ART non produca risultati sempre pienamente soddisfacenti. Infatti, confrontando lo scanpath umano con quello generato da ART in Fig. 11, si osserva che, sebbene la regione delle fissazioni e il target finale coincidano, le fissazioni specifiche presentano delle differenze.

La Fig. 12 mostra lo scanpath generato dal modello DDM-DINO che, pur non essendo perfetto, presenta fissazioni iniziali e finali simili a quelle umane e, in generale, concentra le fissazioni nella medesima area, sebbene con una maggiore dispersione.

Al contrario, ScanDDM ha fornito nuovamente risultati grafici non soddisfacenti, come illustrato in Fig. 13.



Fig. 10: "Sciatori, umano"



Fig. 11: "Sciatori, ART"

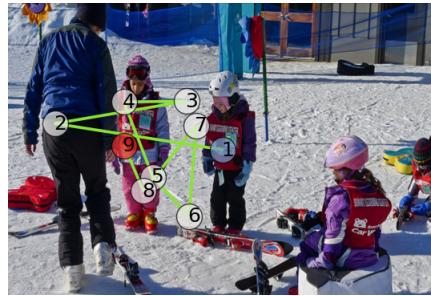


Fig. 12: "Sciatori, DDM-DINO"



Fig. 13: "Sciatori, ScanDDM"

5.1.3 Esempio 3: tavolata

Il terzo e ultimo caso di studio analizzato riguarda il prompt *"person head over glass"*. Lo stimolo visivo è una scena presumibilmente ambientata in un ristorante. In primo piano è visibile una pizza, al centro dell'immagine una donna è ritratta con lo sguardo rivolto verso il volto di un altro individuo, parzialmente occluso da un bicchiere di vetro.

Questo esempio riveste particolare importanza poiché, come si può osservare in Fig. 15, lo scanpath prodotto dal modello ART si discosta marcatamente da quello registrato dall'osservatore umano, come evidenziato in Figura 14. Nello specifico, ART non riesce a individuare

il target dell'attenzione richiesto dal prompt, ovvero la testa che si intravede dietro il bicchiere, concentrandosi invece sul busto della donna.

Anche in questo scenario, lo scanpath generato da ScanDDM, presentato in Fig. 17, devia significativamente dal comportamento visivo umano. Il modello si focalizza inizialmente sul bicchiere per poi dirigere l'attenzione verso il volto di una persona situata nella porzione destra dell'immagine.

Infine, la Fig. 16 illustra il risultato ottenuto impiegando il modello DDM-DINO. In questo caso lo scanpath generato si dimostra notevolmente più allineato con quello umano riportato in Fig. 14. Le fissazioni di DDM-DINO si concentrano inizialmente sulla donna al centro dell'immagine per poi convergere sull'oggetto target della ricerca. Questo specifico esempio dimostra come DDM-DINO possa, in determinate situazioni, superare le prestazioni di ART, che rappresentava l'obiettivo primario di questa ricerca.

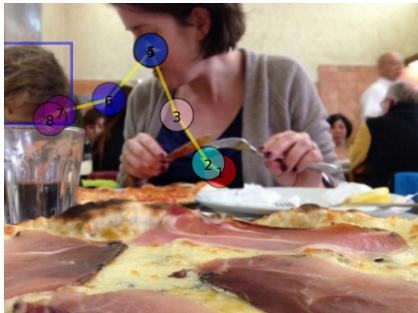


Fig. 14: "Persone, umano"



Fig. 15: "Persone, ART"

5.2 Metriche

L'efficacia del modello proposto, DDM-DINO, in relazione al modello di base ScanDDM e al modello obiettivo ART, è stata quantificata mediante un insieme di metriche di valutazione, la maggior parte delle quali derivate dal framework Fixatons [7]. La selezione delle metriche impiegate è stata guidata dalla loro pertinenza al contesto operativo specifico e dalla loro effettiva computabilità. Di seguito si elencano le metriche adottate per l'analisi comparativa tra scanpath:

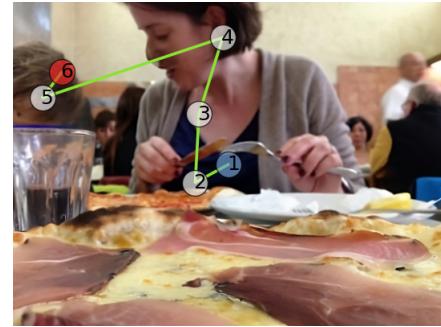


Fig. 16: "Persone, DDM-DINO"

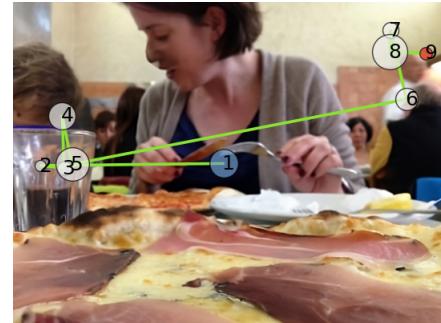


Fig. 17: "Persone, ScanDDM"

- *Euclidean Distance (ED)*: calcola la media delle distanze euclidiene tra le fissazioni corrispondenti di scanpath di lunghezza uguale. A tal proposito, è stata sviluppata una funzione che allinea la lunghezza degli scanpath nel caso avessero un numero di fissazioni differente;
- *String Edit Distance (SED)*: lo stimolo è diviso in zone con associati dei caratteri. Gli scanpath diventano sequenze di caratteri in base a dove cade lo sguardo. La distanza tra due scanpath è misurata in base a quante modifiche servono per rendere uguali le sequenze di caratteri;
- *ScanMatch (SM)*: confronta due sequenze di fissazioni trasformandole in sequenze simboliche basate sulla discretizzazione dello spazio visivo in una griglia. La similarità è calcolata tramite un algoritmo di allineamento di sequenze che considera una matrice di sostituzione basata sulla distanza euclidea tra i bin e una penalità per i gap;
- *Sequence Score (SS)*: analizza la sequenza delle regioni di interesse più rilevanti, identificate tramite clustering, che vengono esplorate. Il confronto si basa sull'ordine di visita di queste regioni, quantificando quanto le traiettorie oculari seguono pattern sequenziali simili nell'esplorazione dello stimolo.

I risultati ottenuti per le diverse metriche sono riportati nella tabella 1. L'analisi di questi valori evidenzia come, in generale, il modello DDM-DINO abbia conseguito prestazioni superiori rispetto al modello di base ScanDDM. Tuttavia, si osserva che in tutte le metriche, ad eccezione della String Edit Distance, i valori prodotti da DDM-DINO rimangono distanti da quelli generati dal

TABLE 1: Metriche

	ED ↓	SED ↓	SM ↑	SS ↑
ScanDDM	0.185	8.804	0.325	0.432
DDM-DINO	0.155	8.635	0.365	0.539
ART	0.102	8.791	0.459	0.585

modello ART, configuratosi come obiettivo prestazionale della presente indagine. Ulteriori considerazioni interpretative in merito ai risultati conseguiti sono presentate nella sezione 6.

6 COMMENTI CONCLUSIVI

L’obiettivo primario del progetto, consistente nell’adattamento del modello ScanDDM al paradigma dell’incremental object referral proposto in ART, è stato conseguito in maniera parziale. Il modello DDM-DINO ha dimostrato una superiorità prestazionale rispetto a ScanDDM, in particolare per quanto concerne i risultati a livello grafico presentati nella sezione 5.1. Gli scanpath generati da DDM-DINO raggiungono l’obiettivo referenziale fornito generalmente, sebbene le fissazioni intermedie non sempre riflettano fedelmente le dinamiche oculari umane.

A livello di metriche, i risultati indicano un miglioramento, limitato seppur presente, rispetto a ScanDDM. Ciò suggerisce che la direzione intrapresa possa essere valida, in quanto si riscontrano comunque esiti positivi, nonostante il raggiungimento della performance di ART rimanga un traguardo significativo. La discrepanza osservata sembra attribuibile alla maggiore complessità infrastrutturale di ART, in contrasto con DDM-DINO che mira unicamente a emulare la logica operativa senza replicarne l’intera architettura complessa. Un’analisi preliminare del codice sorgente conferma la marcata superiorità in termini di complessità strutturale di ART rispetto a ScanDDM, con DDM-DINO che ne eredita la forma più semplice, manifestandosi in un numero considerevolmente inferiore di linee di codice. Un ulteriore aspetto rilevante risiede nel mantenimento della capacità zero-shot in DDM-DINO, caratteristica ereditata da ScanDDM. ART, al contrario, beneficia di un training specifico e approfondito, il che spiega la sua superiore performance nella task richiesta, essendo nativamente progettato e ottimizzato per tale compito, a differenza di DDM-DINO che tenta unicamente di emularne il comportamento senza un addestramento dedicato.

In sintesi, i risultati ottenuti consentono di inferire che l’incremental object referral rappresenta una task di elevata complessità, che esige un’architettura sottostante robusta, una logica di funzionamento sofisticata e un training adeguato, implicando una modifica sostanziale di ScanDDM, non un semplice adattamento. Pertanto, i risultati conseguiti non rappresentano una limitazione definitiva, bensì evidenziano una sfida ancora aperta che potrà stimolare futuri sviluppi.

APPENDIX

ESEMPIO COMPLETO DI OUTPUT

Analogamente a quanto prodotto da ScanDDM, l’output di ciascuna esecuzione di DDM-DINO, come illustrato in Fig. 18, è costituito da molteplici elementi. Nella porzione superiore è riportato il prompt analizzato, comprensivo dei token BOT ed EOT derivanti dal modello ART. Successivamente, sono presentate l’immagine originale, oggetto dell’analisi, e la medesima immagine con sovrapposto lo scanpath generato dal modello. Ogni fissazione è numerata in sequenza e corrisponde a un termine specifico del prompt: la fissazione 1 si riferisce a “BOT”, la 2 a “small”, e così via fino alla 7, corrispondente a “EOT”. Nella sezione inferiore sinistra sono visualizzate le aree dell’immagine che hanno ricevuto una maggiore attenzione da parte degli osservatori durante le simulazioni del DDM, mentre nella porzione destra è mostrata la mappa di salienza generata da Grounding-DINO al termine delle iterazioni.

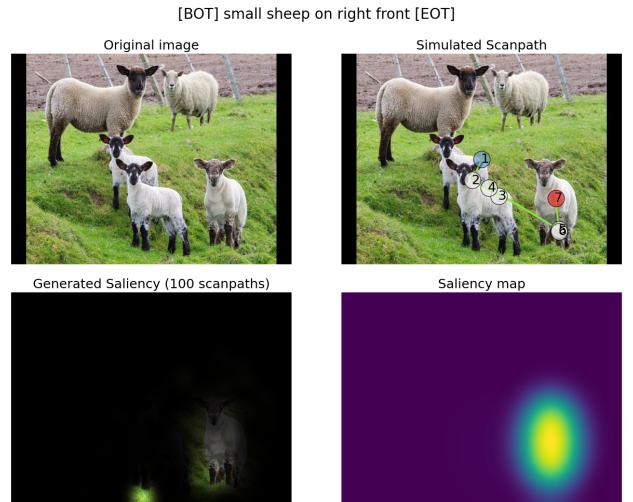


Fig. 18: “Output completo”

REFERENCES

- [1] A. D’Amelio, M. Lucchi, and G. Boccignone, “ScanDDM: Generalised Zero-Shot Neuro-Dynamical Modelling of Goal-Directed Attention,” in *Proceedings of the European Conference on Computer Vision*, 2024.
- [2] S. Mondal, S. Ahn, Z. Yang, N. Balasubramanian, D. Samaras, G. Zelinsky, and M. Hoai, “Look hear: Gaze prediction for speech-directed human attention,” in *European Conference on Computer Vision (ECCV)*, 2024.
- [3] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” *arXiv preprint arXiv:2303.05499*, 2023.
- [4] M. Bucher, T.-H. VU, M. Cord, and P. Pérez, “Zero-shot semantic segmentation,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/0266e33d3f546cb5436a10798e657d97-Paper.pdf
- [5] M. Kümmeler and M. Bethge, “State-of-the-art in human scanpath prediction,” 2021. [Online]. Available: <https://arxiv.org/abs/2102.12239>

- [6] M. Assens, X. G. i Nieto, K. McGuinness, and N. E. O'Connor, "Pathgan: Visual scanpath prediction with generative adversarial networks," 2018. [Online]. Available: <https://arxiv.org/abs/1809.00567>
- [7] D. Zanca, V. Serchi, P. Piu, F. Rosini, and A. Rufa, "Fixatons: A collection of human fixations datasets and metrics for scanpath similarity," *CoRR*, vol. abs/1802.02534, 2018. [Online]. Available: <http://arxiv.org/abs/1802.02534>