ACADEMIA ROMÂNĂ

Școala de Studii Avansate a Academiei Române

Institutul de Matematiică ”Simion Stoilow”

REZUMATUL TEZEI DE DOCTOAT

De la vedere computațională la
procesarea limbajului natural

CONDUCĂTOR DE DOCTORAT:
Prof. Dr. Marius Leordeanu

DOCTORAND:
Ing. Mihai-Dan Mașala

2025

Cuprins

ABSTRACT

Această teză propune un cadru nou pentru gândirea și prelucrarea datelor cu modalități multiple, printr-o reprezentare explicită, inspirată de om, bazată pe grafuri. GEST - Graph of Events in Space and Time (Graf de eveniment în spațiu și timp) reprezintă o reprezentare universală, construită în primul rând pentru a reprezenta poveștile ca coloană vertebrală a comunicării umane, care se află între mai multe modalități, la fel cum creierul uman procesează, integrează și raționează asupra mai multor modalități.

Utilizând GEST, valorificăm tehnicile existente de învățare automată pentru a demonstra puterea sa expresivă și pentru a aborda sarcini bine stabilite, cum ar fi generarea text-video și descrierea video, prin această nouă perspectivă. Natura explicită a GEST adaugă un nivel de explicabilitate atât la sarcinile de tip video-to-text, cât și text-to-video. Putem înțelege de ce o anumită acțiune a fost menționată în descrierea textuală, putem înțelege relația sa spațială și temporală cu alte acțiuni și o putem evidenția cu acuratețe în cadrul înregistrării video. Cadrul propus în această teză, nu numai că atinge performanțe mai bune, dar reprezintă, de asemenea, un pas spre inteligența artificială explicabilă.

În general, această teză introduce o nouă reprezentare între viziune și limbaj, redefinind modul standard de gândire și rezolvare a sarcinilor la intersecția acestor două domenii. Folosind metode de învățare profundă, dovedim puterea reprezentării propuse, rezolvăm două sarcini end-to-end existente și validăm eficacitatea acesteia prin evaluări umane și automate extinse.

Cuvinte cheie – *Vedere computațională, Procesarea limbajului natural, Vedere și limbaj, Generare de video, Descrierea video bogată semantic*

Capitolul 1

INTRODUCERE

Oamenii posedă o capacitate extraordinară de a procesa, integra, raționa și crea în mod transparent prin intermediul mai multor modalități. Această capacitate este esențială pentru viața noastră de zi cu zi și este atât de adânc înrădăcinată încât o luăm adesea de bună. De exemplu, relatarea unei scene la care am fost martori unei alte persoane pare lipsită de efort, naturală și un aspect esențial al interacțiunii sociale. Devine chiar o nevoie, o dorință. În plus, această abilitate, alături de multe altele, este fundamentală pentru societatea noastră, servind drept pilon al comunicării și al experiențelor comune.

La nivel personal, cunoașterea este un proces remarcabil de complicat și de multifațetat, deoarece multiplele reprezentări nu există independent, ci sunt profund interconectate. Mai degrabă decât să funcționeze independent, multitudinea de reprezentări interacționează dinamic pentru a ne modela percepția. Cercetările au arătat că cititul activează nu numai regiunile creierului care procesează limbajul, ci și zonele asociate cu percepția vizuală, ajungând până la activarea sistemului nervos [1].

Această teză își propune să investigheze acest fenomen complex printr-o perspectivă computațională. În mod specific, ea urmărește să exploreze modul în care progresele recente în vederea computațională și prelucrarea limbajului natural pot face lumină în acest proces complex de cunoaștere multimodală. Această cercetare dorește să

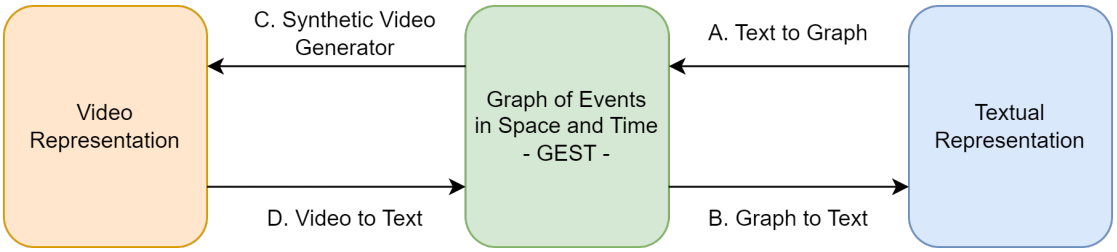

FIGURA 1.1: Prezentare generală funcțională a cadrului propus, centrată pe GEST. GEST reprezintă componenta centrală, permițând tranziții fără întreruperi între diferite forme. De exemplu, tranziția de la text la video se face prin etapele A și C, în timp ce transformarea de la video la text se poate face prin etapele D și B.

înțeleagă mai bine mecanismul care stă la baza cunoașterii multimodale din perspectivă computațională. **Principalele contribuții** ale acestei teze sunt rezumate mai jos:

Reprezentare. Unul dintre obiectivele principale ale acestei teze este de a explora modul în care mintea umană integrează în mod natural reprezentări multiple într-un spațiu de reprezentare unificat. Apare o întrebare fundamentală: există cu adevărat un astfel de spațiu sau este doar o construcție teoretică? Cercetările în domeniul reprezentării multimodale sugerează că creierul uman poate codifica diferite reprezentări (de exemplu, vederea și limbajul) într-un spațiu latent comun, cu existența neuronilor multimodali [2]. În locul unei reprezentări implicite, numerice, propunem o reprezentare explicită, explicabilă, bazată pe grafuri, sub forma GEST. Pe lângă introducerea acestei reprezentări noi, demonstrăm, de asemenea, puterea expresivă a GEST, capacitatea și eficiența sa de a surprinde și de a modela elementele complexe ale poveștilor, validând alegerile noastre de proiectare.

Integrarea multiplelor reprezentări. Scopul nostru este de a construi o reprezentare care integrează și agregă diferite reprezentări, la fel cum creierul uman sintetizează informațiile prin gânduri. Trecerea de la o modalitate la alta (sau chiar la aceeași) se va face prin, sau utilizând GEST. Această teză se concentrează pe două dintre cele mai răspândite reprezentări: vederea și limbajul (a se vedea Figura 1.1). În mod specific, ne propunem să investigăm conversiile și tranzițiile ca atare: text către GEST, GEST către text, video către GEST și GEST către video.

Generare text-video prin GEST. În primul rând, investigăm metode semi-automate și automate de construire a **GEST din text**, urmate de un modul **GEST-to-video** care

exploatează natura explicită a GEST pentru a construi videoclipuri precise și complexe. Cu aceste două module, abordăm sarcina **text-la-video**: pornind de la texte, construim GEST asociate și apoi videoclipuri. Prin intermediul evaluării umane și automate, demonstrăm că videoclipurile noastre oferă o reprezentare precisă și eficientă a textului.

De la video la text prin GEST. Propunem o abordare algoritmică multitask care exploatează sarcini de viziune pe calculator bine stabilite (de exemplu, segmentarea semantică, detectarea acțiunilor) pentru a extrage automat GEST-uri din videoclipuri (**video-to-GEST**). Concentrându-ne pe acțiuni și actori, extragem grafuri care urmăresc cu exactitate ceea ce se întâmplă în videoclip. Pentru sarcina **GEST-la-text**, propunem o abordare în două etape care generează mai întâi un proto-limbaj, urmată de utilizarea LLM bazat pe text pentru rafinarea acestuia. Proto-limbajul reprezintă o descriere text completă care conține toate acțiunile și actorii prezenți în GEST, descriere text care este construită algoritmic pe baza sortării și grupării evenimentelor în spațiu și timp, urmată de reguli bazate pe gramatică pentru a descrie fiecare grup. Realizăm o evaluare aprofundată a seturilor de date și a metodelor existente pentru această sarcină, dezvăluind lipsa resurselor existente. Arătăm că chiar și seturile de date care sunt considerate în mod tradițional complexe sau foarte lungi pot fi descrise printr-o simplă legendă. Pe lângă expunerea diferențelor fundamentale dintre seturile de date video, atât evaluările umane, cât și cele automate arată că, în cazul seturilor de date potrivite pentru descrierea video bogată, abordarea noastră propusă depășește cu mult alte metode de ultimă generație. În plus, arătăm că extinderea abordării noastre cu sarcini de nivel superior (de exemplu, descrierea cadrelor) îmbunătățește și mai mult calitatea descrierilor generate, cu rezultate de ultimă oră în toate seturile de date și metrici.

GEST și VLLMs. În cele din urmă, ridicăm câteva întrebări și efectuăm câteva experimente inițiale cu privire la cum și dacă natura explicită și explicabilă a GEST poate fi combinată cu puternicele Visual Large Language Models pentru a spori modelele existente în scopul de a obține rezultate fundamentate, controlabile și explicabile. Pentru moment, ne concentrăm pe sarcina de a adapta LLM pentru limba română, de a construi resurse pentru formare și evaluare și de a adăuga capacități multimodale sub formă de input vizual.

Capitolul 2

GRAFURI DE EVENIMENTE IN SPATIU SI TIMP - GEST

În esență, un GEST este un mijloc de a reprezenta povești. Ne concentrăm pe modelarea poveștilor, deoarece acestea sunt principala modalitate de exprimare a ideilor, sentimentelor, faptelor, percepțiilor, întâmplărilor din lumea reală sau fantastică. Poveștile sunt o componentă esențială în teatru, în cinematografie sub formă de storyboard-uri și sunt, de asemenea, o parte integrantă în relatarea, comunicarea și predarea evenimentelor istorice. Poveștile sunt universale: o viață este o poveste, un vis este o poveste, un singur eveniment este o poveste. Evenimentele atomice creează povești complexe în același mod în care părțile mici formează un obiect într-o imagine sau cum cuvintele formează o propoziție. Prin urmare, în modelarea poveștilor, distingem interacțiunile în spațiu și timp ca fiind componenta centrală. În general, schimbările în spațiu și timp conduc la noțiunea de evenimente și interacțiuni. În mod similar cu modul în care schimbările într-o imagine (gradienții imaginii) pot reprezenta margini, schimbările spațiu-timp (la diferite niveluri de abstractizare) reprezintă evenimente. În consecință, evenimentele în

Acest capitol este baza pe lucrarea - *(ArXiv, 2023)* ***Mihai Masala****, Nicolae Cudlenco, Traian Rebedea, and Marius Leordeanu. GEST: the Graph of Events in Space and Time as a Common Representation between Vision and Language. arXiv preprint arXiv:2305.1294. 2023. [3]*

spațiu și timp ar putea fi detectabile, repetabile și discriminatorii. Interacțiunile dintre evenimentele din spațiu și timp schimbă starea actuală a lumii, pot declanșa sau provoca alte evenimente și, la rândul lor, pot provoca alte schimbări. Prin urmare, utilizăm aceste evenimente și interacțiunile lor în spațiu și timp drept componentă fundamentală a GEST. În esență, o margine conectează două evenimente în spațiu și timp. Această conexiune poate fi temporală (de exemplu, după, între timp), logică (de exemplu, și, sau) sau spațială (de exemplu, deasupra), dar nu este limitată la acestea. Deoarece un nod în GEST poate reprezenta, de asemenea, obiecte fizice (de exemplu, „Casa există pentru această perioadă de timp"), conexiunile grafului pot reprezenta orice relație potențială între două obiecte sau două evenimente: evenimentul „casă" a fost implicat în evenimentul „organizarea unei reuniuni la casa respectivă". Prin urmare, o muchie poate reprezenta și un eveniment prin el însuși.

În plus, cadrul nostru GEST este un pas înainte față de abordarea clasică Subiect-Verb-Obiect (SVO). În cazul nostru, subiectul devine un eveniment (chiar dacă vorbim despre evenimente de tipul „există", acestea sunt totuși evenimente) și, de asemenea, obiectul devine un eveniment. Un eveniment este compus din obiecte, iar orice eveniment necesită interacțiune între obiecte și lume. Deoarece în formularea noastră obiectele sunt evenimente, orice interacțiune (și deci orice margine) devine în sine un eveniment. Acest lucru permite o reprezentare ierarhică și recursivă în GEST. Modelele clasice reprezintă interacțiuni de la obiect la obiect, pe care GEST le poate reprezenta la fel de ușor. În plus, putem trece la nivelul următor, modelând hiper-evenimente, colapsând astfel de interacțiuni la un singur nod, generând un proces recursiv infinit în care nodurile se extind și se colapsă în evenimente. Chiar și evenimentele simple pot fi explicate printr-un GEST, deoarece toate evenimentele pot fi împărțite, la un nivel de detaliu suficient, în evenimente mai simple și în interacțiunile lor (de exemplu, "deschid ușa" devine un GEST complex dacă descriem în detaliu mișcările mâinii și componentele mecanice implicate). În același timp, orice graf GEST poate fi văzut ca un eveniment unic de la o scară semantică și spațio-temporală superioară (de exemplu, „o revoluție politică" ar putea fi atât un graf GEST, cât și un eveniment unic). Colapsarea grafurilor în noduri ($Event \Leftarrow GEST$) sau extinderea nodurilor în grafuri ($GEST \Leftarrow Event$) oferă

GEST posibilitatea de a avea mai multe niveluri de profunzime, după cum este necesar pentru povestiri vizuale și lingvistice complexe.

După cum s-a menționat anterior, pentru fiecare eveniment codificăm în principal tipul de acțiune, entitățile implicate, locul și intervalul de timp în care are loc un eveniment. În mod esențial, în GEST atât evenimentele explicite (de exemplu, acțiunile), cât și cele implicite (de exemplu, existența) sunt reprezentate utilizând aceleași reguli. Un exemplu complet de GEST poate fi găsit în Figura 2.1. Relația cauzală dintre două evenimente (și anume, E2 deoarece E10) poate fi exprimată ca un singur eveniment. În principiu, orice GEST poate deveni un eveniment într-un GEST de nivel superior și viceversa, orice eveniment poate fi extins într-un GEST mai detaliat.

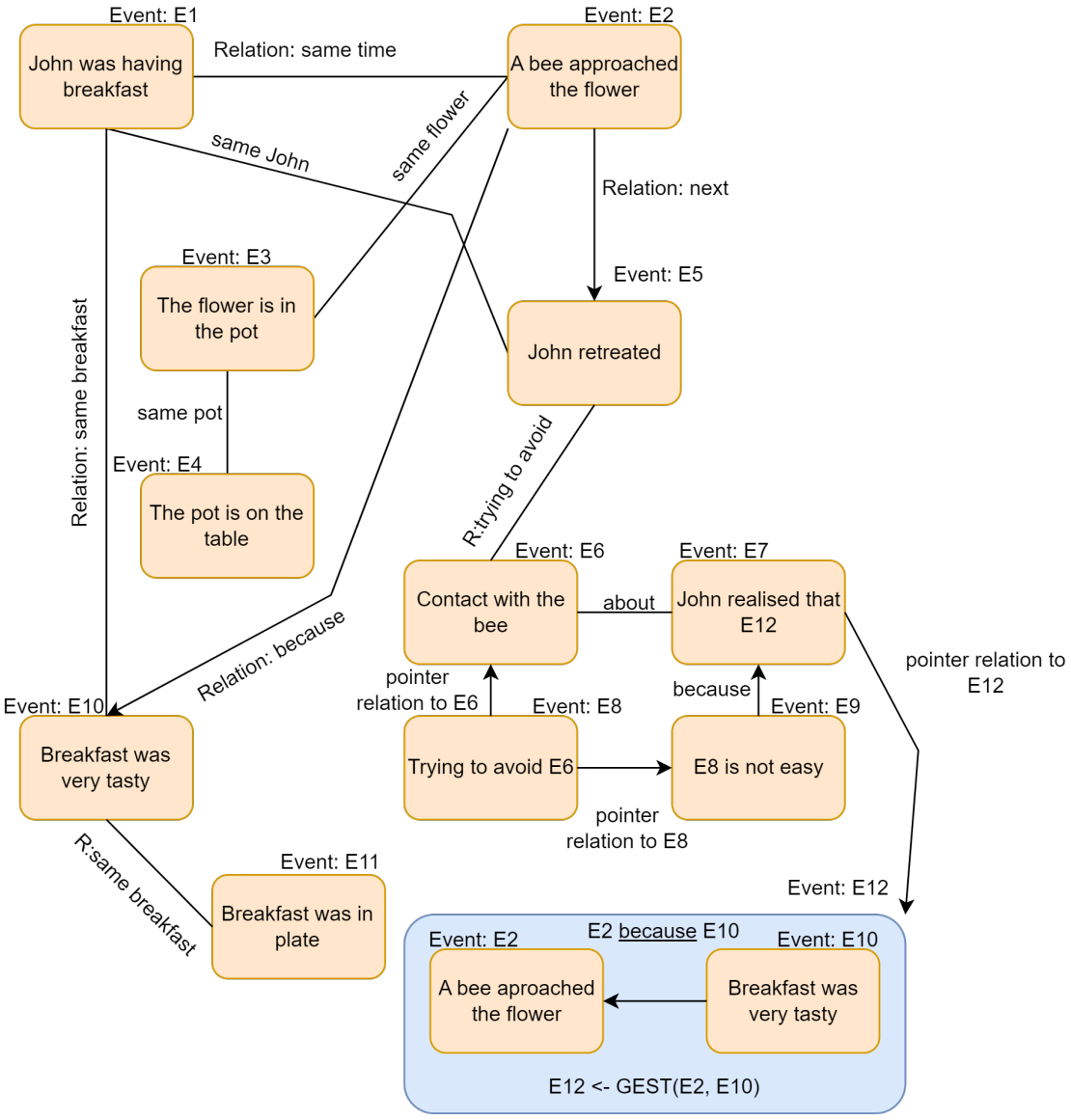

FIGURA 2.1: GEST grafic pentru explicarea textului următor: *"John was having breakfast when a bee approached the flower in the pot on the table. Then he pulled back trying to avoid contact with the bee but he realized that it was not an easy attempt because she actually came because of the tasty food on his plate"*.

Capitolul 3

DE LA TEXT LA GEST: PUTEREA GEST

Grafurile de evenimente în spațiu și timp oferă o reprezentare comună și semnificativă pentru mai multe modalități. În acest capitol, ne concentrăm pe construirea grafurilor de evenimente în spațiu și timp pornind de la reprezentări textuale. Traducerea textelor în aceste noi spații de reprezentare creează noi oportunități. Pentru a dovedi puterea de reprezentare a GEST și pentru a valida abordarea noastră, abordăm sarcina similarității textelor dintr-un nou punct de vedere: în loc să comparăm direct texte, aducem textele în spațiul GEST și efectuăm comparații prin intermediul metricii de potrivire a grafurilor în acest spațiu nou. Arătăm că, pentru sarcina de a detecta dacă două texte provin din același videoclip, abordarea noastră bazată pe GEST depășește metricile clasice de generare a textelor și poate, de asemenea, să sporească performanța metricilor de ultimă generație, puternic antrenate.

Acest capitol este bazat pe lucrarea - *(ICCVW, 2023)* ***Mihai Masala****, Nicolae Cudlenco, Traian Rebedea, and Marius Leordeanu. Explaining vision and language through graphs of events in space and time. In Workshop on Closing the Loop Between Vision and Language at the IEEE/CVF International Conference on Computer Vision Workshops. 2023. [4]*

3.1 Construirea GEST din text

GEST din text este necesar pentru formare și evaluare. Observăm că construirea reprezentării GEST din text nu este o sarcină trivială și ne propunem să automatizăm acest proces. Cu toate acestea, pentru a obține un GEST corect din text este încă necesară intervenția umană. Din fiecare propoziție, dorim să extragem informații precum tipul de acțiuni, entitățile implicate, locațiile și momentele acțiunilor, precum și relațiile dintre acestea. Toate acestea sunt informații extrase prin analiza arborelui de dependență (extras automat[1]) al fiecărei propoziții individuale folosind un set de reguli realizate manual (urmate de corecție umană dacă este necesar). Contextul (de exemplu, inferența locației) și ordinea evenimentelor sunt de asemenea injectate în grafic pentru a obține GEST complet al unei povești. În timp ce pentru corpusul bAbI [5] toate entitățile (de exemplu, actori, obiecte) sunt unice pentru fiecare poveste (de exemplu, un singur actor cu numele John în fiecare poveste), în cazul Videos-to-Paragraphs [6] acest lucru nu este întotdeauna cazul. Prin urmare, trebuie să intervenim manual și să setăm referințele adecvate (construiți și legați numărul adecvat de noduri), deoarece diferite entități sunt menționate cu același nume în SVO-uri (de exemplu, „om", „desk"). Pentru a găsi și adnota cu precizie aceste cazuri, parcurgem manual fiecare pereche de SVO și poveste și le verificăm semantic validitatea.

3.2 Similaritate pentru GEST

Pentru a compara GEST-urile, evaluăm două metode de potrivire grafică, o abordare clasică, Potrivirea spectrală (SM) [7] și o abordare modernă bazată pe învățarea profundă, Potrivirea grafică neuronală (NGM) [8]. SM este o metodă rapidă, robustă și precisă care utilizează vectorul propriu principal al unei matrice de afinitate, în timp ce NGM folosește mai multe rețele neuronale care învață și transformă matricea de afinitate în graficul de asociere, care este încorporat și utilizat ca intrare pentru un clasificator de vârfuri. Pentru următoarele experimente, luăm în considerare toate perechile pozitive din setul de testare Video-to-Paragraphs (un total de 67 în cazul nostru) și 174 de perechi

[1] `https://spacy.io/models/en#en_core_web_lg` accesat ultima dată pe 19 ianuarie 2023

negative eșantionate aleatoriu din același set de test. Atât pentru algoritmii SM, cât și pentru NGM, matricea de afinitate este construită folosind atât funcții de similaritate la nivel de nod, cât și la nivel de muchii care exploatează embedding-uri pre-antrenate. Folosim embedding-uri GloVe [9] pre-antrenate de dimensiunea 300, pentru a măsura similitudinea la fiecare nivel (de exemplu, acțiune, entități) pentru noduri. Pentru a compara două muchii, integrăm asemănarea la nivel de nod (de la nodurile care sunt conectate la muchiile particulare) cu similaritatea la nivel de margine (adică asemănarea dintre tipul de muchii). În esență, două noduri sunt la fel de similare ca și acțiunile și entitățile lor, în timp ce asemănarea a două muchii este dată prin înmulțirea tipului de muchie (de exemplu, următorul, între timp) asemănarea cu asemănarea nodurilor corespunzătoare.

3.3 Rezultate și Discuții

Rezultatele din Tabelul 3.1 atestă puterea reprezentării noastre propuse: potrivirea grafurilor în spațiul GEST depășește toate metricile clasice de generare de text (adică BLEU@4, METEOR și ROUGE) și chiar și valorile moderne bazate pe Transformatori pre-antrenați, cum ar fi BERTScore.

Method	**Correlation**	**Accuracy**	**Fisher**	**AUC**
BLEU@4	24.45	75.52	0.2816	52.65
METEOR	58.48	84.23	1.1209	73.90
ROUGE	51.11	83.40	0.7164	68.92
SPICE	59.42	84.65	1.0374	74.43
BERTScore	57.39	85.89	1.0660	<u>77.93</u>
GEST Spectral	<u>61.70</u>	84.65	<u>1.2009</u>	75.47
GEST Neural	60.93	<u>86.31</u>	0.9770	76.75

TABELA 3.1: Rezultate care compară puterea de reprezentare GEST (adevărul de bază) cu valorile comune de generare de text aplicate pe poveștile din setul de testare Videos-to-Paragraphs. Arătăm cu **bold** cea mai bună valoare pentru fiecare măsură, iar cu <u>underline</u> cea mai bună valoare.

Testul inițial arată puterea de reprezentare a GEST, dar nu testați capacitatea acestei reprezentări de a fi combinată cu una puternic antrenată. Testăm această capacitate arătând că GEST poate stimula o metrică de ultimă generație, puternic antrenată, chiar și

atunci când le combinăm pe cele două în cel mai simplu mod liniar. Pornind de la textul inițial al poveștii, învățăm să transformăm automat povestea în GEST (prin reglarea fină a unui model GPT3, text-curie-001[2]), și apoi prin comparație, prin compararea, folosind un scor de potrivire GEST, folosind un grafic EST corespunzător. GEST-uri generate. Un al doilea scor BLEURT între povești se obține ca și înainte. Învățăm apoi, pe setul de antrenament, cum să combinăm liniar cele două partituri, pentru a separa cel mai bine textele aceleiași povești de textele diferitelor povești. Aplicam aceeasi procedura tuturor metricilor clasice, pentru a evalua beneficiul adus de GEST fata de alte metode. În Tabelul 3.2 arătăm rezultatele BLEURT (sus), cele ale altor valori combinate cu BLEURT folosind aceeași abordare de regresie liniară (din mijloc) și rezultatele GEST (jos), folosind cele două metode de potrivire grafică (SM și NGM). Este important de reținut că, în combinație cu alte valori, BLEURT nu se îmbunătățește întotdeauna, dar atunci când este combinat cu GEST se îmbunătățește întotdeauna și cu cea mai mare marjă.

Method	**Correlation**	**Accuracy**	**Fisher**	**AUC**
BLEURT	70.93	90.04	2.0280	88.02
+BLEU@4	70.93	90.04	2.0274	88.04
+METEOR	71.20	89.63	2.0659	87.62
+ROUGE	70.76	90.04	1.9973	87.71
+SPICE	<u>71.94</u>	88.80	<u>2.0808</u>	87.71
+BERTScore	71.11	89.63	2.0089	87.25
+GEST Spectral	**72.89**	**90.87**	**2.2086**	**89.80**
+GEST Neural	71.91	<u>90.46</u>	2.0537	<u>88.58</u>

TABELA 3.2: Rezultate care compară puterea BLEURT împreună cu valorile comune de generare de text și GEST (învățat), aplicate pe poveștile din setul de testare Video-to-Paragraphs. Notațiile sunt aceleași ca în Table 3.1.

Experimentele noastre demonstrează puterea GEST: noul său spațiu și măsurarea asociată de potrivire a grafurilor pot fi utilizate în mod eficient, cu costuri minime de instruire, pentru a crește performanța valorilor de ultimă generație existente. Chiar și cu date foarte limitate, experimentele noastre arată că GEST este mai mult decât potrivit pentru a recrea povestea de bază, într-un spațiu care permite comparații foarte fiabile și corelate umane.

[2] `https://platform.openai.com/docs/models/gpt-3` accesat ultima dată pe 8 mai 2023

Capitolul 4

GENERAREA DE VIDEO-URI DIN GEST

Inteligența artificială face progrese mari astăzi și începe să reducă decalajul dintre viziune și limbaj. Cu toate acestea, suntem încă departe de a înțelege, explica și controla în mod explicit conținutul vizual din perspectivă lingvistică, deoarece încă ne lipsește o reprezentare comună explicabilă între cele două domenii. Ne îndreptăm atenția spre generarea de videoclipuri din reprezentările GEST, pentru a stabili GEST ca o reprezentare comună între viziune și limbaj.

4.1 De la text la video prin GEST

Pentru a finaliza legătura dintre GEST și lumea vizuală, introducem motorul poveștilor vizuale. Bazat pe jocul GTA San Andreas cu Multi Theft Auto (MTA)[1], care interacționează

Acest capitol este bazat pe lucrarea - *(ICCVW, 2023)* ***Mihai Masala****, Nicolae Cudlenco, Traian Rebedea, and Marius Leordeanu. Explaining vision and language through graphs of events in space and time. In Workshop on Closing the Loop Between Vision and Language at the IEEE/CVF International Conference on Computer Vision Workshops. 2023. [4]*

[1] `https://multitheftauto.com/` accesat ultima dată pe 25 iulie 2023

cu mecanica jocului, folosim locațiile, obiectele și animațiile preexistente din joc și ne concentrăm asupra evenimentelor care au loc în și în jurul unei case. Motorul are control deplin în mediul virtual și poate, prin urmare, să profite din plin de natura structurată și explicabilă a GEST. Remarcăm că întregul proces este complet automatizat și nu necesită intervenție umană, permițând astfel generarea video la scară. Sistemul ia ca intrare un GEST și, pe baza acestuia, generează mai multe videoclipuri valide. Împreună cu modulul text-to-GEST menționat anterior, am închis bucla și am construit un sistem care este capabil să genereze videoclipuri din text. În continuare, generăm un set de 25 de videoclipuri complexe a câte 2-3 minute fiecare, cu până la 15 activități diferite, mult mai mari decât cele folosite în literatura actuală. Chiar dacă setul este mic, este foarte dificil, așa că îl folosim pentru a valida abordarea noastră.

În continuare, prezentăm atât evaluări umane, cât și evaluări automate ale videoclipurilor noastre generate de GEST, în comparație cu modelele recente text-to-video [10, 11].

4.2 Evaluare

Invităm adnotatorii umani să evalueze videoclipurile în ceea ce privește conținutul semantic fără text introdus, pe o scară de la 1 la 10 și să aleagă cel mai bun videoclip pentru fiecare text introdus. În total, am colectat 111 adnotări. Pentru fiecare text, utilizatorilor li se prezintă cu textul original cele trei videoclipuri generate și o baterie de întrebări privind calitatea generală a fiecărui videoclip.

Videoclipurile generate de noi au fost evaluate cu un scor mediu de 7.93, în timp ce videoclipurile generate folosind Text2VideoZero [11] și CogVideo [10] au un scor mediu de 5.55 și respectiv 4.03, așa cum este prezentat în Figura 4.1. În peste 85% din cazuri, evaluatorii umani au ales videoclipurile noastre ca fiind cele mai bune per ansamblu, cu aproximativ 11% din cazuri în care Text2VideoZero a fost considerat cel mai bun.

Folosim VALOR [12], o metodă recentă de generare video-în-text, și măsurăm cât de bine se potrivește textul generat înapoi din videoclipurile generate cu textele inițiale de intrare. VALOR este antrenat și testat separat pentru fiecare tip de metodă de generare

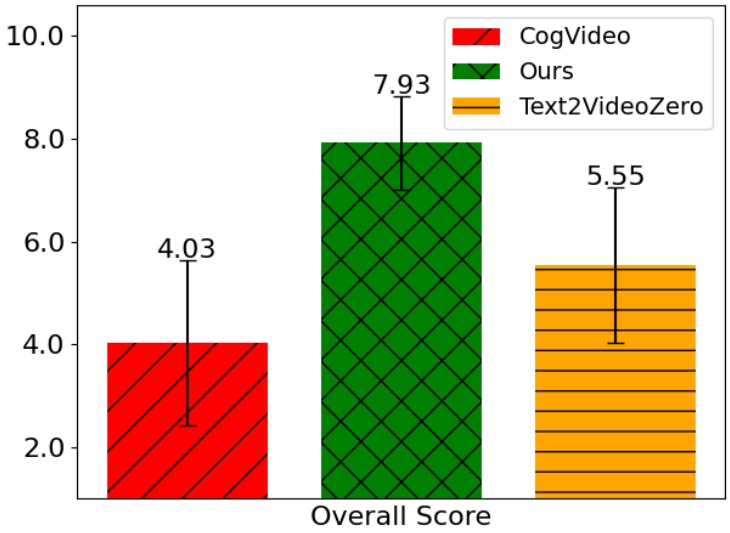

FIGURA 4.1: Scoruri (de la 1 la 10) date de evaluatorii umani.

Metric	GEST	CogVideo	Text2VideoZero
Bleu@4	9.84	8.16	**10.02**
Meteor	**14.16**	13.48	13.96
ROUGE	**35.40**	32.72	34.87
SPICE	**20.04**	19.54	19.43
CIDEr	**34.12**	33.16	33.65
BERTScore	**19.37**	13.09	15.02
BLEURT	**39.44**	37.55	38.40

TABELA 4.1: Rezultate pentru conversia video-la-text. Afișăm cu **bold** cea mai bună valoare pentru fiecare metrică.

video, folosind validare încrucișată în 5 fold-uri, de la zero, pe parcursul a 3 rulări, cu rezultatele mediate (prezentate în Tabelul 4.1). Aceste experimente corespund evaluării umane, menținând aceeași ierarhie între metode și demonstrând că videoclipurile generate de GEST pot păstra mai bine conținutul semantic al textului original de intrare. Aceasta dovedește că un model viziune-limbaj explicit și complet explicabil, sub forma unui grafic de evenimente în spațiu și timp, ar putea oferi în practică o modalitate mai bună de a explica și controla conținutul semantic – aducând astfel o valoare complementară în contextul modelelor AI realiste (dar nu neapărat adevărate).

Capitolul 5

EXTRAGEREA GEST DIN VIDEOCLIPURI

În capitolele anterioare am explorat puterea reprezentării GEST (prin sarcina text-to-GEST) și modul în care GEST poate fi utilizat pentru a genera videoclipuri (adică, GEST-to-video). În capitolele următoare explorăm piesele care lipsesc, și anume cum să construiți un GEST dintr-un videoclip (acest capitol) și cum să generați o descriere bogată dintr-un anumit GEST (capitolul următor). O prezentare generală a abordării noastre propuse este prezentată în Figura 5.1.

Pentru a construi o reprezentare explicită a unui videoclip, incorporăm multiple sarcini — în principal din domeniul viziunii computerizate (detecția acțiunii, segmentarea semantică, estimarea adâncimii). Pentru fiecare cadru dintr-un videoclip dat, extragem mai întâi aceste informații, urmate de un pas de potrivire și agregare. Ieșirea detectorului de acțiuni include, pentru fiecare acțiune, o casetă de delimitare a persoanei care efectuează acțiunea împreună cu numele acțiunii și un scor de încredere. Pornind de la

Acest capitol este baza pe lucrarea - *(ArXiv, 2025)* ***Mihai Masala****, and Marius Leordeanu Towards Zero-Shot & Explainable Video Description by Reasoning over Graphs of Events in Space and Time. arXiv preprint arXiv:2501.08460. 2025. [13]*

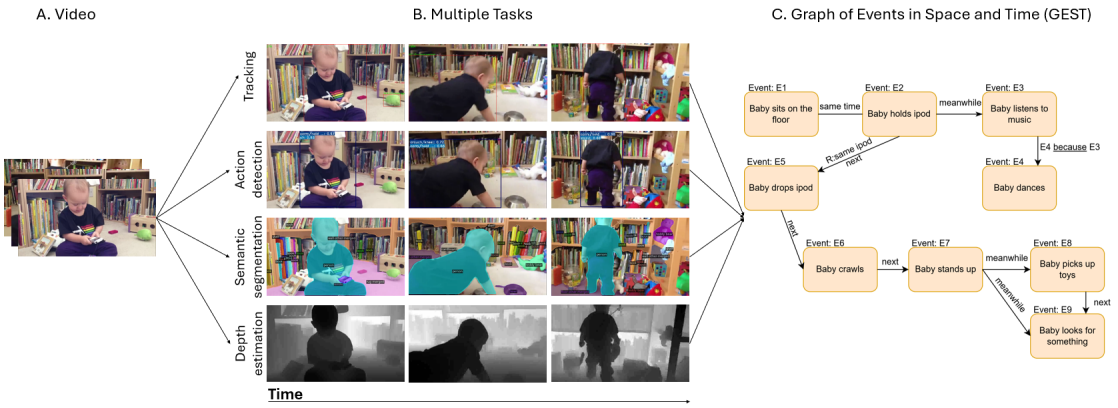

FIGURA 5.1: O prezentare generală a abordării noastre. Pornind de la un videoclip brut, realizăm detectarea și urmărirea obiectelor, detectarea acțiunilor, segmentarea semantică și estimarea adâncimii. Agregăm aceste informații pentru a construi Graful corespunzător al Evenimentelor în Spațiu și Timp.

această casetă de delimitare, urmărim să adunăm toate obiectele din vecinătatea persoanei, obiecte cu care actorul ar putea interacționa. Această listă de obiecte este filtrată pe baza diferenței de adâncime dintre persoană și obiect pentru a păstra doar cele relevante. În continuare, aplicăm un filtru de acțiuni bazat pe cadru, păstrând doar acțiunile cu cel mai mare scor de încredere, implementând de asemenea un mecanism de vot cu o fereastră de 11 cadre. Următorul pas este să agregăm și să procesăm informația la nivel de cadru într-o formă globală la nivel de videoclip. Rezolvăm inconsistențele pe termen scurt prin unificarea a două id-uri de persoane dacă apar apropiate în timp (mai puțin de 10 cadre) și se suprapun suficient (intersecție peste uniune mai mare de 0.6), în timp ce inconsistențele pe termen lung (cum ar fi atunci când o persoană iese din cadru și reintră într-o etapă și locație diferită) sunt rezolvate printr-un modul semantic, bazat pe aparență, care utilizează histograme HSV. După aceea, agregăm acțiunile care apar în cadre consecutive (permitem câteva cadre lipsă) și salvăm cadrul de început și sfârșit, obiectele implicate (uniunea) și casetele de delimitare.

Ultimul pas în toată aceast proces constă în construirea de relații spațio-temporale între evenimente: construim perechi de evenimente și dacă acestea îndeplinesc anumite criterii le legăm în spațiu (distanța euclidiană măsurată a centroizilor) și timp (următorul, același timp sau între timp).

Capitolul 6

CONSTRUIREA UNOR DESCRIERI BOGATE SI ANCORATE DIN GEST

Așa cum am abordat anterior sarcina de la video la GEST, este destul de natural să discutăm despre sarcina de a descrie videoclipul în limbaj natural. Totuși, ne îndepărtăm de sarcina de subtitrare video, deoarece, în general, descrierile lingvistice din subtitrarea video sunt foarte simple. În acest capitol, propunem un teren comun între viziune și limbaj, bazat pe GEST, într-un mod explicabil și programatic, pentru a oferi o soluție la problema de lungă durată a descrierii videoclipurilor în limbaj natural bogat.

6.1 Metodă

Traducerea unui GEST într-o descriere coeră, bogată și naturală nu este o sarcină simplă, cu posibilități multiple. În această lucrare adoptăm o abordare în două etape care valorifică puterea LLM-urilor existente bazate pe text pentru a construi descrieri naturale.

Părți din acest capitol sunt bazate pe lucrarea - *(ArXiv, 2025)* ***Mihai Masala****, and Marius Leordeanu Towards Zero-Shot & Explainable Video Description by Reasoning over Graphs of Events in Space and Time. arXiv preprint arXiv:2501.08460. 2025. [13]*

Scopul primului pas al abordării noastre este de a converti graficul într-un sunet, dar poate aspru în jurul marginilor, forma textuală, o descriere inițială pe care o numim proto-limbaj. Pentru a obține o descriere mai asemănătoare omului, folosim LLM-urile existente, hrănindu-le cu acest proto-limbaj și îndemnând cu scopul de a rescrie textul pentru a-l face să sune mai natural.

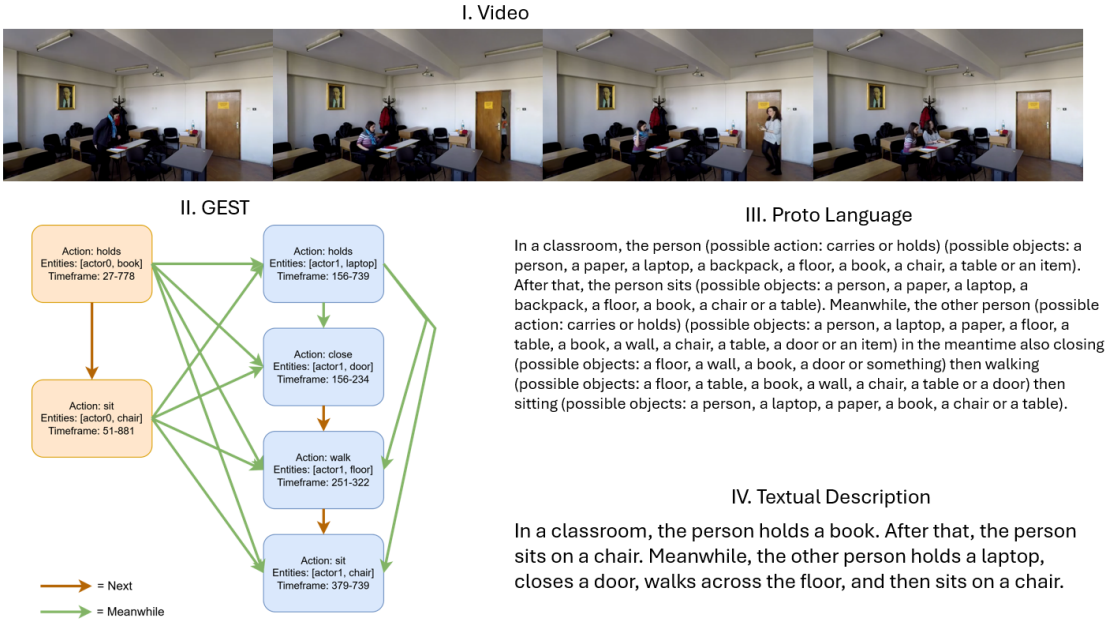

FIGURA 6.1: Un exemplu complet al conductei noastre propuse. Pornind de la video-clip, construim automat GEST-ul asociat (vezi Capitolul 5). Din acest grafic, construim proto-limbajul care este apoi alimentat unui LLM care generează descrierea textuală finală.

Construirea proto-limbajului. Primul pas în acest proces implică o sortare temporală a graficului (după cadrul de început al fiecărui eveniment; asemănător unei sortări topologice). Abordarea noastră reunește acțiunile ordonate cronologic în grupuri de acțiuni de nivel superior de către actori. Fiecare astfel de grup este apoi descris în text, prin descrierea fiecărui eveniment folosind o gramatică simplă și luând în considerare relațiile spațiale și temporale intra-grup și inter-grup. Când descriem un eveniment, listăm toate obiectele posibile și lăsăm LLM să aleagă obiectele care sunt cel mai probabil să apară în contextul dat, cu puterea de a alege un nou obiect care nu este prezent în listă sau să nu aleagă deloc un obiect. În plus, permitem LLM să schimbe numele unei acțiuni sau să șteargă complet o acțiune și entitățile asociate acesteia dacă nu se potrivește contextului. Un exemplu complet este prezentat în Figura 6.1.

Method	Text Metrics	Humans	LLM Jury
VidIL [14]	4 (13.24)	3 (2.84)	4 (3.21)
GIT2 [15]	3 (13.61)	2 (2.73)	2 (2.71)
mPLUG-2 [16]	5 (12.14)	3 (2.84)	3 (2.75)
PDVC [17]	2 (14.18)	5 (4.88)	5 (4.82)
GEST	1 (15.05)	1 (1.71)	1 (1.51)

TABELA 6.1: Ordonarea absolută (cel mai bun este 1, cel mai rău este 5, urmată de valoarea absolută a fiecărei valori) a metodelor în cadrul diferitelor evaluări (metrici de generare de text, evaluare umană și un juriu format din LLM-uri) pentru setul de date Video-to-Paragraphs.

Method	Average	VtP (489)	COIN (318)	WebVid (443)	VidOR (478)	VidVRD (164)
VidIL [14]	3.21	4.00	2.82	2.93	3.21	3.11
GIT2 [15]	3.58	3.79	3.39	3.61	3.57	3.55
mPLUG-2 [16]	3.53	3.85	3.11	3.63	3.63	3.44
PDVC [17]	5.50	5.77	5.15	5.65	5.39	5.53
GEST	3.16	1.96	3.79	3.33	3.16	3.55
GEST + VidIL	**2.02**	**1.64**	**2.74**	**1.86**	**2.03**	**1.84**

TABELA 6.2: Clasamentul mediu (cel mai bun este 1, cel mai rău este 6) selectat de juriul LLM. VtP - Videoclipuri în paragrafe. **Bold** marchează cel mai bun rezultat din fiecare categorie, în timp ce underline marchează al doilea cel mai bun. Metoda de cea mai bună performanță, așa cum este evaluată folosind abordarea LLM-as-a-Jury, este din nou combinația dintre GEST și VidIL.

6.2 Results and Discussions

Rezultatele sunt prezentate în Tabelul 6.1 și Tabelul 6.2. Pe setul de date Videos-to-Paragraphs, GEST depășește clar celelalte metode, în toate cele trei direcții de evaluare. Îmbunătățirea metodei noastre cu un set mai divers de acțiuni și obiecte sau, echivalent, ancorarea setului bogat de intrări VidIL în acțiuni clare și concrete – ambele obținute simultan prin combinarea GEST și VidIL – conduce la descrieri mai bune. Astfel de descrieri sunt ancorate, conțin mai puține halucinații și descriu mai bine video-ul sursă.

În acest capitol introducem o metodă nouă pentru a descrie cu acuratețe Grafuri de Evenimente în Spațiu și Timp folosind un limbaj natural bogat. Combinând acest modul cu modulul Video-to-GEST prezentat în capitolul anterior, putem aborda descrieri bogate, de tip poveste, ale videoclipurilor.

Capitolul 7

O PRIVIRE CATRE VIITOR. GEST SI (V)LLM-URI

În epoca actuală a învățării automate, în care există o abundență de date, atât reale, cât și sintetice, neutilizarea unor modele puternic antrenate ar fi o oportunitate ratată. Chiar dacă, așa cum am arătat în capitolele anterioare, calitatea datelor disponibile nu este întotdeauna optimă, modelele de ultimă generație extrag în continuare informații valoroase din acestea și sunt capabile de performanțe aproape umane. În acest context, cum se aliniază GEST cu peisajul actual? În loc să concuram cu modele foarte bine pregătite și deja performante, credem că GEST poate servi ca o soluție complementară. Și am arătat deja că această abordare poate funcționa atât atunci când combinăm GEST cu o metrică de generare de text de ultimă generație (vezi Capitolul 3), cât și când combinăm GEST cu alte metode pentru sarcina text-to-video (vezi Capitolul 6). În mod similar, GEST ar putea fi utilizat ca sursă de fundamentare în fluxurile de lucru text-to-video și video-to-text. Metodele text-to-image au început deja să integreze

Părți din acest capitol sunt bazate pe lucrarea - *(EMNLP, 2024)* ***Mihai Masala****, Denis C Ilie-Ablachim, Alexandru Dima, Dragos Corlatescu, Miruna Zavelca, Ovio Olaru, Simina Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. "Vorbești Românește?" A Recipe to Train Powerful Romanian LLMs with English Instructions. In Findings of the Association for Computational Linguistics: EMNLP 2024. [18]*

informații suplimentare care fundamentează și controlează imaginea generată.Lian et al. [19] folosesc un LLM disponibil comercial pentru a extrage entități din promptul text și a construi un layout fundamentat al imaginii. Apoi, un model de difuzie stabil ghidat de layout-ul generat este utilizat pentru a crea imaginea finală. În implementarea originală, atât generatorul de layout, cât și generatorul de imagine sunt metode generale, disponibile comercial, fără a se efectua antrenamente suplimentare sau ajustări fine. Chiar dacă aceasta este sarcina inversă, abordarea este foarte similară cu metoda noastră propusă de video-to-text. Ambele împărtășesc o reprezentare intermediară (adică, layout și proto-limbaj) și ambele folosesc metode și tehnici de vârf disponibile comercial.Zhang et al. [20] au lucrat la adăugarea controlului (de exemplu, postură, adâncime, segmentare) la modelele text-to-image, în timp ce Ashual and Wolf [21] controlează imaginea generată printr-un grafic al scenei.

7.1 GEST și (V)LLM-uri

În acest capitol, căutăm să investigăm cum poate fi utilizat GEST împreună cu metode de ultimă generație text-to-video, în principal modele vizuale de limbaj mari. Unde și cum ar putea fi integrat GEST? Este suficient să adăugați pur și simplu GEST în prompt; modelele ar trebui antrenate sau ajustate cu GEST? Care este cel mai bun mod de a integra GEST, ca grafic, ca proto-limbaj sau ca descriere finală?

Modelele de limbă mari multimodale (MLLM) se bazează pe succesul LLM-urilor (numai text), care servesc ca o evoluție naturală care integrează mai multe modalități. Modelele Visual Large Language (VLLM) adaugă capacități de viziune la LLM-urile existente, permițându-le să proceseze, să înțeleagă și să raționeze asupra intrării vizuale (adică imagini și videoclipuri). Dezvoltarea VLLM a urmat o traiectorie similară cu cea a LLM-urilor, Alayrac et al. [22] fiind primul care a explorat învățarea în context pentru sarcini de viziune și limbaj, urmată de paradigma de reglare fină a instrucțiunilor vizuale [23] care a preluat domeniul.

Pentru a investiga dacă și cum putem integra GEST în VLLM-urile existente, realizăm un set inițial de experimente în care, pe lângă cadrele video, adăugăm fie proto-limbajul

Method	Avg	B@4	M	R	C	S	BS	BT
GPT-4o	19.31	**2.85**	**16.55**	**24.21**	1.19	**15.70**	23.90	**50.79**
w/ Proto-Language	**19.56**	2.78	14.31	23.71	**5.22**	15.10	**27.42**	48.39
w/ Description	17.55	1.89	14.77	21.78	1.83	12.47	21.87	48.25

TABELA 7.1: Rezultatele Videos-to-Paragraphs atunci când se folosesc subtitrările annotate de oameni ca referință de adevăr. **Bold** marchează cel mai bun rezultat din fiecare categorie, în timp ce underline marchează al doilea cel mai bun. B@4 reprezintă Bleu@4, M pentru Meteor, R pentru ROUGE-L, C pentru CIDEr, S pentru SPICE, BS pentru BERTScore și BT pentru BLEURT.

(împreună cu instrucțiunile „de asamblare”), fie descrierea finală. Rezultatele din Tabelul 7.1 arată că integrarea unei astfel de informații direct în VLLM-urile existente nu este simplă. Adăugarea proto-limbajului la input-ul video crește abia calitatea descrierilor generate, depășind doar linia de bază pe două metrici (CIDEr și BERTScore). Folosirea descrierii finale produce chiar rezultate mai slabe, performând semnificativ mai slab comparativ cu linia de bază. Conjecturăm că acest fenomen provine din procedura de antrenament a VLLM-urilor, deoarece acestea nu sunt de obicei condiționate pe mai multe modalități pentru sarcini de generare de descriere/video. Prin urmare, credem că cele mai eficiente metode pentru integrarea informațiilor GEST în VLLM-urile moderne sunt în timpul fazei de antrenament, fie prin pre-antrenament, fine-tuning sau în faza de aliniere.

Din cauza unei combinații de lipsă de modele, date, resurse hardware și know-how în pregătirea unor astfel de modele mari la începutul acestei călătorii, decidem să începem cu o sarcină mai simplă, aproape de jucărie. Acest lucru reduce mult resursele necesare, atât la nivel hardware, cât și la nivel de date, ne permite să înțelegem mai bine procesul, avantajele și limitările unor astfel de modele. Prin urmare, alegem sarcina de a adapta VLLM la limba română, atât ca modalitate de a ne permite o înțelegere mai profundă a acestor modele, cât și o modalitate de a da înapoi comunității.

Deoarece orice VLLM începe cu un codificator vizual pre-antrenat și un LLM pre-antrenat, primul pas logic este evaluarea eficienței modelelor existente în limba română. După cum vom arăta, în special cu modelele anterioare (singurele disponibile la începutul acestei căi de cercetare), performanța lor în limba română este semnificativ în urmă față

de cunoașterea limbii engleze într-o gamă largă de sarcini, unele modele nefiind capabile nici măcar să răspundă la o întrebare în limba română. Acest lucru este de așteptat deoarece aceste modele sunt concentrate în principal pe limba engleză, aproape 90% din datele utilizate pentru instruire fiind în limba engleză, în timp ce datele românești au reprezentat doar 0,03%.

Pentru adaptarea LLM-urilor și VLLM-urilor existente în limba română, recurgem la traducerea atât a seturilor de date de instruire, cât și a benchmark-urilor de evaluare din surse engleze de înaltă calitate. Adăugăm, de asemenea, sarcini în aval nativ românesc și elaborăm un nou etalon de referință care evaluează cunoștințele culturale (despre România și cultura română) ale LLM.

Model	**Avg**	**ARC**	**MMLU**	**Wino**	**HS**	**G8k**	**TQA**
			Llama2				
Llama2-7b	37.04	36.05	33.66	57.56	48.00	4.75	42.22
RoLlama2-7b-Base	38.03	37.95	27.22	59.29	57.22	2.53	44.00
Llama2-7b-chat	36.84	37.03	33.81	55.87	45.36	4.90	44.09
RoLlama2-7b-Instruct	**44.50**	44.73	40.39	63.67	59.12	13.29	45.78
			Mistral				
Mistral-7B-v0.1	45.02	42.99	47.16	60.77	54.19	16.20	48.80
Mistral-7B-Instruct-v0.2	47.40	46.29	47.01	58.78	54.27	13.47	**64.59**
RoMistral-7b-Instruct	**52.91**	**52.27**	49.33	**70.03**	**62.88**	32.42	50.51
			Llama3				
Llama3-8B	44.55	38.05	48.33	59.94	53.48	20.04	47.44
Llama3-8B-Instruct	50.62	43.69	52.04	59.33	53.19	43.87	51.59
RoLlama3-8b-Instruct	**52.21**	47.95	53.50	66.06	59.72	40.16	45.90
			Llama3.1				
Llama3.1-8B	37.29	33.25	36.35	58.80	42.65	3.59	49.03
Llama3.1-8B-Instruct	49.87	42.86	53.73	59.71	56.82	35.56	50.54
RoLlama3.1-8b-Instruct	**53.03**	47.69	**54.57**	65.85	59.94	**44.30**	45.82
			Gemma				
gemma-7b	50.04	47.22	53.18	61.46	60.32	30.48	47.59
gemma1.1-7b-it	41.39	40.05	47.12	54.62	47.10	9.73	49.75
RoGemma-7b-Instruct	**50.48**	52.02	52.37	66.97	56.34	25.98	49.18
			Other models				
Okapi-Ro	35.64	37.90	27.29	55.51	48.19	0.83	44.15
aya-23-8B	**45.81**	43.89	45.96	60.50	60.52	16.81	47.16

TABELA 7.2: Comparison între RoLLMs și alte LLM-uri pe versiunile românești ale benchmark-urilor academice (abrevieri: HS - HellaSwag, Wino - Winogrande, G8k - GSM8k, TQA - TruthfulQA). **Bold** denotă cel mai bun în fiecare categorie (medie) și în total (fiecare benchmark). Observăm îmbunătățirea generală a versiunii românești pe toate familiile de modele.

În Tabel 7.2 prezentăm rezultate privind reperele academice pentru RoLLM. Observăm o îmbunătățire consistentă a RoLLMS adaptate în comparație cu omologii lor inițiali. Pentru RoVLLM, rezultatele sunt prezentate în Tabel 7.3. La fel ca și în cazul RoLLM-urilor, observăm o îmbunătățire consistentă a VLLM-urilor românești în sarcinile luate în considerare.

Model (%)	**AVG**	**MMBench**	**MMStar**	**MMMU**
Baselines				
LLaVA-Mistral-7b	39.93	58.06	32.73	29.00
LLaVA-Llama3-8b	41.31	60.40	32.53	31.00
Visual-Textual Alignment				
LLaVA-Llama3-8b	26.64	27.50	26.53	25.89
LLaVA-RoLlama3-8b	26.51	25.56	25.87	27.11
Visual Instruction Tuning				
LLaVA-Llama3-8b	44.16	66.76	36.60	29.11
LLaVA-Llama3-8b	44.54	67.25	34.80	31.56
LLaVA-RoLlama3-8b	**45.72**	**68.46**	**36.93**	**31.78**

TABELA 7.3: Performanța VLLM pe benchmark-uri românești. *Italic* marchează modelul construit de la zero: adaptorul inițializat aleatoriu, LLM de bază așa cum este. Liniile de bază (fără cursive) au fost deja aliniate și ajustate pe datele în limba engleză. Rețineți că modelele marcate cu *italic* în secțiunea inferioară sunt versiuni ajustate ale modelelor prezentate în secțiunea din mijloc (modele aliniate). **Bold** marchează cel mai bun rezultat pentru fiecare benchmark.

În acest capitol, am introdus primele LLM-uri și VLLM-uri open-source specializate pentru limba română. Evaluările noastre arată rezultate promițătoare, depășind soluțiile existente pe mai multe benchmarcuri. Pentru ambele LLM-uri și VLLM-uri, prezentăm o rețetă generală de antrenament și evaluare, inclusiv resursele, rețetă pe care ne așteptăm să funcționeze și pe alte arhitecturi. La fel ca și cu RoLLM-urile, prezentăm o rețetă generală de antrenament și evaluare, inclusiv resursele, rețetă pe care ne așteptăm să funcționeze și pe alte arhitecturi, mai mari și mai puternice.

Direcțiile viitoare sunt împărtășite cu RoLLMs, deoarece folosim în principal aceeași abordare: creșterea numărului și calității seturilor de date, prin validarea traducerilor și colectarea de instrucțiuni și date de preferință umană. În plus, pentru RoVLLMS, în această etapă nu am realizat alinierea preferințelor umane pentru modelele actuale, o etapă pe care am arătat-o ca fiind critică pentru performanța LLM.

Capitolul 8

CONCLUZII SI DIRECTII VIITOARE

Această teză a introdus un cadru și o reprezentare inedită pentru a legături între viziune și limbaj. Am oferit atât baze teoretice, cât și metode și rezultate practice puternice, care pot fi extinse la alte modalități. În plus, nu am abordat toate sarcinile de la intersecție și viziune și limbaj (de exemplu, răspunsul la întrebări vizuale) și suntem siguri că mai multe aplicații interesante ale acestui cadru există și vor fi studiate (de exemplu, raționamentul în spațiul GEST). De-a lungul cercetării noastre, am dezvoltat abordări noi, construind resurse valoroase pe parcurs. Este important de menționat că aceste lucrări reprezintă doar primii pași în definirea și utilizarea GEST la întregul său potențial și numeroase căi de cercetare sunt încă deschise.

Mai jos, oferim un rezumat al contribuțiilor pe care le aduce această teză:

Reprezentare. Introducem o reprezentare nouă, bazată pe grafuri, sub forma Grafului de Evenimente în Spațiu și Timp. Creăm GEST cu scopul principal de a construi o reprezentare universală a secvențelor de evenimente, a poveștilor. Considerăm această reprezentare apropiată de modul în care poveștile sunt reprezentate și utilizate în creierul uman. Oferim intuiția din spatele GEST, definiții formale și exemple ale capacității și universalității sale și arătăm cum putem construi GEST-uri din text.

Generarea textului în video. După ce am generat deja GEST-uri din text, următoarea sarcină a fost generarea GEST-to-video. Folosind resursele existente ale unui motor de joc (MTA), am conceput o procedură pentru a construi o poveste vizuală dintr-un GEST. Natura explicită a GEST combinată cu abordarea algoritmică garantează acuratețea videoclipului generat, indiferent de lungimea și complexitatea acestuia.

Descrierea detaliată a videoclipului. Am explorat și sarcina inversă, de la video la text. Exploatând natura bogată a GEST, ne-am concentrat pe sarcina de descriere a videoclipului de tip poveste, spre deosebire de sarcina mai simplă de generare a legendelor pentru videoclipuri. Folosind o abordare multi-task, extragem din fiecare videoclip un Grafic al Evenimentelor în Spațiu și Timp (Capitolul 5). A doua etapă a procesului video-to-text, anume GEST-to-text, este abordată folosind o combinație de sortare a grafurilor (în spațiu și timp), reguli bazate pe gramatică și LLM-uri pentru a descrie cu acuratețe în text evenimentele prezente în GEST.

Combinarea GEST cu VLLM-uri. Inspirat de experimentele anterioare, în care am arătat că GEST este complementar soluțiilor de ultimă oră pentru metricile de generare a textului și descrierile video, am început să investigăm dacă și cum poate fi combinat GEST cu modelele vizuale mari de limbaj (VLLM-uri), care sunt în mare parte inexplicabile, dar foarte puternice. Experimentele inițiale au arătat că integrarea unei modalități suplimentare sub forma GEST (chiar și în formă textuală) nu este chiar simplă, așa că am început o misiune de a înțelege mai bine cum să antrenăm LLM-uri și cum să adăugăm cel mai bine reprezentări suplimentare.

8.1 Agenda viitoare de cercetare

Considerăm că călătoria noastră de până acum reprezintă doar începutul GEST și aplicațiile sale. Această teză tocmai a pus bazele și a identificat o parte din provocările către o punte explicabilă între viziune și limbaj. Sperăm că acest lucru inspiră și deschide calea pentru ca alți cercetători să depășească limitările muncii noastre și să deschidă noi orizonturi pentru sarcinile de viziune și limbaj.

În continuare, schițăm câteva direcții de cercetare care reprezintă extensii naturale ale muncii noastre, care nu au fost încă investigate din cauza limitărilor de timp și resurse.

- **Creșterea numărului de sarcini.** Deși am propus soluții pentru două sarcini de viziune și limbaj, descrierea video și generarea text-video, nu am discutat despre cum poate fi utilizat GEST pentru a rezolva alte sarcini și intersecția dintre viziune și limbaj, cum ar fi întrebările vizuale cu răspuns (VQA) sau implicațiile vizuale (VE) [24]. În VQA, scopul este de a răspunde la o întrebare în limbaj natural despre o imagine sau un video. Pentru această sarcină, GEST ar putea fi esențial, deoarece un posibil flux de lucru ar putea include un proces multi-pași care convertește video-ul într-un GEST și întrebarea naturală într-o interogare, care poate fi utilizată pentru a interoga graful. În general, în această teză am folosit caracteristica GEST pentru ancorare, însă nu ne-am concentrat în mod special pe aplicații care implică raționamente asupra GEST sau chiar modificarea grafului (de exemplu, pentru generarea de exemple adversariale).

- **Integrarea GEST în VLLMs și Modele de Difuzie.** Așa cum s-a arătat deja în această teză, informațiile codificate în GEST pot completa soluțiile existente de vârf. Adăugarea de ancorare prin GEST fie în VLLMs, fie în modelele de difuzie ar putea crește semnificativ complexitatea și acuratețea textelor sau videoclipurilor generate. Am construit deja resurse valoroase, inclusiv tuple de video-GEST și liste de descrieri textuale ordonate, care pot fi utilizate direct (de exemplu, ca set de date bazat pe preferințele umane pentru descrierea video).

- **Îmbunătățirea și extinderea resurselor existente.** Conectându-ne la direcția anterioară, am generat deja videoclipuri pornind de la text. Folosind aceste date și aplicând tehnici existente, cum ar fi segmentarea semantică pentru a extrage cutii de legătura ale persoanelor, am putea genera aranjamente pentru un model de generare vizuală (de exemplu, Modele de Difuzie). Extinderea suplimentară a resurselor generate, atât în număr, cât și în calitate, poate fi utilizată pentru a îmbunătăți și mai mult metodele existente. În cele din urmă, fiecare dintre metodele dezvoltate, implementate și evaluate în această teză are potențialul de a

fi îmbunătățită, în special având în vedere progresele remarcabile din LLM-uri și VLLM-uri.

8.2 Observații de final

În această eră în care inteligența artificială a devenit omniprezentă, asistenții AI funcționează acum fără probleme pe dispozitive portabile, capabile să integreze și să genereze videoclipuri, vorbire și text. În zilele noastre, AI este aplicată în aproape toate domeniile – medicină, drept, educație, sport, politică și multe altele. Deși aceste modele demonstrează capacități remarcabile într-o gamă largă de sarcini și cu siguranță vor continua să evolueze, credem că avem responsabilitatea de a studia și de a investi în siguranța și explicabilitatea AI.

Dincolo de atingerea și chiar depășirea performanței umane a reperelor, o performanță impresionantă în sine, este imperativ să ne străduim și să interpretăm și să înțelegem procesele de luare a deciziilor ale unor astfel de modele AI. Iar explicabilitatea vine într-o varietate de moduri, inclusiv abordări precum expunerea raționamentului [25], perturbarea intrării [26] sau tehnici bazate pe gradient [27]. Mai mult, într-o lume în care AI este mai mult ca probabil utilizată în domenii cu miză mare, inclusiv aplicații militare, siguranța AI devine primordială.

Sperăm că această activitate contribuie și inspiră comunitatea de cercetare, ghidând dezvoltarea unor sisteme AI corecte, sigure, imparțiale, explicabile și de încredere.

Bibliografie

[1] Stephen M Kosslyn, Giorgio Ganis, and William L Thompson. Neural foundations of imagery. *Nature reviews neuroscience*, 2(9):635–642, 2001.

[2] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435 (7045):1102–1107, 2005.

[3] Mihai Masala, Nicolae Cudlenco, Traian Rebedea, and Marius Leordeanu. Gest: the graph of events in space and time as a common representation between vision and language. *arXiv preprint arXiv:2305.12940*, 2023.

[4] Mihai Masala, Nicolae Cudlenco, Traian Rebedea, and Marius Leordeanu. Explaining vision and language through graphs of events in space and time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2826–2831, 2023.

[5] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[6] Simion-Vlad Bogolin, Ioana Croitoru, and Marius Leordeanu. A hierarchical approach to vision-based language generation: from simple sentences to complex natural language. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2436–2447, 2020.

[7] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Tenth IEEE International Conference on*

Computer Vision (ICCV'05) Volume 1, volume 2, pages 1482–1489 Vol. 2, 2005. doi: 10.1109/ICCV.2005.20.

[8] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Neural graph matching network: Learning lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[10] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[11] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

[12] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.

[13] Mihai Masala and Marius Leordeanu. Towards zero-shot & explainable video description by reasoning over graphs of events in space and time. *arXiv preprint arXiv:2501.08460*, 2025.

[14] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497, 2022.

[15] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[16] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. *ArXiv*, abs/2302.00402, 2023.

[17] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6847–6857, 2021.

[18] Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. "vorbești românește?" a recipe to train powerful Romanian LLMs with English instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.681.

[19] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[21] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4561–4569, 2019.

[22] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al.

Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[24] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[25] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL `https://arxiv.org/abs/2501.12948`.

[26] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, 2015.

[27] Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. *arXiv preprint arXiv:2305.15853*, 2023.