ROMANIAN ACADEMY

School of Advanced Studies of the Romanian Academy

„Simion Stoilow" Institute of Mathematics

PHD THESIS SUMMARY

Bridging the Gap Between Computer Vision and Natural Language Processing

THESIS ADVISOR:

Prof. Dr. Marius Leordeanu

PHD STUDENT:

Eng. Mihai-Dan Mașala

2025

Contents

ABSTRACT

This thesis proposes a novel framework for thinking and working with multiple modalities, through an explicit, human-inspired, graph-based representation. GEST - Graph of Events in Space and Time represents a universal representation, built primarily for representing stories as the backbone of human communication, stands between multiple modalities just as the human brain processes, integrates and reasons over multiple modalities.

Using GEST, we leverage existing machine learning techniques to show its expressive power and address well established tasks, such as text-to-video generation and video description, through this new lens. The explicit nature of GEST adds a layer of explainability to both video-to-text and text-to-video end-to-end tasks. We can understand why a particular action was mentioned in the textual description, understand its spatial and temporal relation with other actions, and accurately highlight it within video. Our proposed framework not only achieves better performance but also represents a step towards explainable AI.

Overall, this thesis introduces a novel representation between vision and language, redefining the standard way of thinking and solving tasks at the intersection of these two domains. Using deep learning methods, we prove the power of the proposed representation, solve two existing end-to-end task, and validate its effectiveness through extensive human and automatic evaluation.

Keywords – *Computer Vision, Natural Language Processing, Vision and Language, Video generation, Rich Video Description*

Chapter 1

INTRODUCTION

Humans possess an extraordinary capacity to seamlessly process, integrate, reason and create across multiple modalities. This ability is central to our everyday life and so deeply embedded that we often take it for granted. For instance, relating a scene we have witnessed to another person feels effortless, natural and an essential aspect of social interaction. It even becomes a need, a desire. Moreover, this, along many other skills, is fundamental to our society, serving as pillar of communication and shared experiences.

On a personal level, cognition is a remarkably intricate and multifaceted process, as multiple modalities do not exist independently but are deeply intertwined. Rather than functioning independently, the multitude of modalities interact dynamically to shape our perception. Research found that reading activates not only language-processing regions of the brain but also areas associated with visual perception, going all they way to activating the nervous system [1].

This thesis aims to investigate this complex phenomenon through a computational lens. Specifically, it seeks to explore how recent advancements in computer vision (CV) and natural language processing (NLP) can shed some light into this complex multi-modal cognition process. This research wants to better understand the underlying mechanism

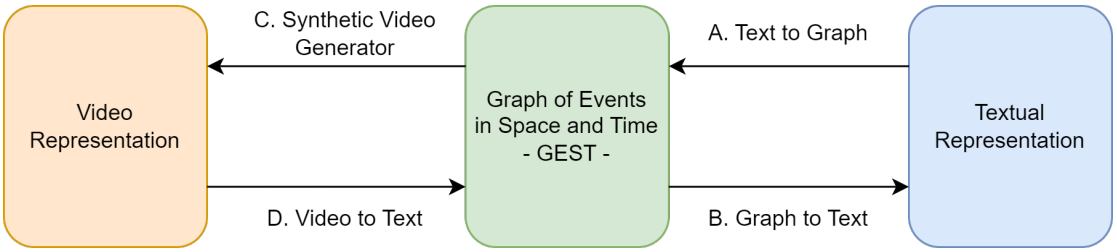

FIGURE 1.1: Functional overview of the proposed framework, centered around GEST. GEST represent the central component, allowing for seamless transitions between different forms. For example the transition from text to video is done via steps A and C, while the transformation from video to text can be done via steps D and B.

of multi-modal cognition from a computation perspective. The **main contributions** this thesis makes are summarized below:

Representation. One of the primary objectives of this thesis is to explore how the human mind naturally integrates multiple modalities into a unified representation space. A fundamental question arises: does such a space truly exist, or is it merely a theoretical construct? Research in multimodal representation suggests that the human brain may encode different modalities (e.g., vision and language) into a common latent space with the existence of multimodal neurons [2]. Instead of an implicit, obfuscated, numerical representation we propose an explicit, explainable, graph-based representation in the form of Graph of Events in Space and Time (GEST). Besides introducing this novel representation, we also prove GEST's expressive power, its ability and effectiveness to capture and model intricate elements of stories, validating our design choices.

Multimodality integration. Our goal is to build a representation that integrates and aggregates different modalities, just as the human brain synthesizes information through thoughts. Switching from one modality to another (or even the same one) will go through, or utilizing GEST. This thesis focuses on two of the most widespread modalities: vision and language (see Figure 1.1). Specifically, we aim to investigate conversions and transitions as such: text to GEST, GEST to text, video to GEST, and GEST to video. By addressing and solving these tasks effectively, we lay the foundation for a functionality similar to that of the human brain.

Text-to-Video generation through GEST. First, we investigate semi-automated and automated methods of building **GEST from text** followed by a **GEST-to-video** pipeline

harnesses the explicit nature of GEST to build accurate, grounded and complex videos. With these two modules, we tackle the **text-to-video** task: starting from texts, we build associated GEST and then videos. Through both human and automatic evaluation, we demonstrate that our videos offer a more accurate and effective representation of the text, compared to other state-of-the-art approaches.

Video-to-Text description through GEST. We propose an algorithmic multi-task approach that harnesses well established computer vision tasks (e.g., semantic segmentation, action detection) to automatically extract GESTs from videos (**video-to-GEST**). By focusing on actions and actors, we extract grounded graphs that accurately follow what happens in the video. For **GEST-to-text** task, we propose a two stage approach that first generates a proto-language, followed by using text-based LLM for refining it. The proto-language represents a complete text description that contains all the actions and actors present in the GEST, text description that is built algorithmically based on sorting and grouping events in space and time, followed by grammar-based rules to describe each group. We perform an thorough evaluation of existing datasets and methods for this task, revealing the lack of existing resources. We show that even datasets that are traditionally considered complex or very long, could be described by a simple caption. Besides exposing fundamental differences between video dataset, both human and automated evaluations show that on datasets suited for rich video description, our proposed approach outperforms by far other state-of-the-art methods. Furthermore, we show that extending our approach with higher-level tasks (e.g., frame captioning) further improves the quality of the generate descriptions with state-of-the-art results across datasets and metrics.

GEST and VLLMs. Finally, we pose some questions and perform some initial experiments on how and if the explicit and explainable nature of GEST can be combined with powerful Visual Large Language Models to augment existing models in order to obtain grounded, controllable and explainable outputs. For the time being we focus on the task adapting LLMs for Romanian, building resources for training and evaluation, and adding multimodal capabilities in the form of visual input.

Chapter 2

GRAPH OF EVENTS IN SPACE AND TIME – GEST

Fundamentally, a GEST is a means of representing stories. We focus on modeling stories as they are the main way of expressing ideas, sentiments, facts, perceptions, real-world or fantasy happenings. Stories are an essential component in theater, cinema in the form of storyboards and are also an integral part in relating, communicating and teaching historical events. Stories are universal: a life is a story, a dream is a story, a single event is a story. Atomic events create intricate stories in the same way that small parts form an object in a picture, or how words form a sentence. Therefore, in modeling stories, we distinguish interactions in space and time as the central component. In general, changes in space and time lead to the notion of events and interactions. Similarly to how changes in an image (image gradients) might represent edges, space-time changes (at different levels of abstraction) represent events. Accordingly, events in space and time could be detectable, repeatable and discriminative. Interactions between events in space and time change the current state of the world, can trigger or cause other

This chapter is based on the paper - *(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]*

events and in turn cause other changes. Therefore, we use these events and their interactions in space and time as the fundamental component of GEST. Fundamentally, an edge connects two events in space and time. This connection can be, but is not limited to temporal (e.g. after, meanwhile), logical (e.g. and, or) or spatial (e.g. on top of). Since a node in GEST can also represent physical objects (e.g, "The house exists for this period of time") the graph connections can represent any potential relation between two objects or two events: the event "house" was involved in event: "holding a meeting at that house". Therefore, an edge can also represent an event by itself.

Furthermore, our GEST framework is a step-up from the classic Subject-Verb-Object (SVO) approach. In our case, the Subject becomes an event (even if we are talking about events of type "exists", they are still events) and also the Object becomes an event. An event is composed by objects, and any event requires interaction between objects and the world. As in our formulation objects are events, any interaction (and so any edge) becomes in itself an event. This allows a hierarchical and recursive representation in GEST. Classic models represent object to object interactions, that GEST can easily represent as well. Moreover, we can go to the next level, modeling hyper-events, collapsing such interactions to a single node, generating an infinite recursive process in which nodes expand and collapse into events. Even simple events can be explained by a GEST, since all events can be broken, at a sufficient level of detail, into simpler ones and their interactions (e.g. "I open the door" becomes a complex GEST if we describe in detail the movements of the hand and the mechanical components involved). At the same time, any GEST graph could be seen as a single event from a higher semantic and spatio-temporal scale (e.g. "a political revolution" could be both a GEST graph and a single event). Collapsing graphs into nodes ($Event \Leftarrow GEST$) or expanding nodes into graphs ($GEST \Leftarrow Event$), gives GEST the possibility to have many levels of depth, as needed for complex visual and linguistic stories.

As previously mentioned, for each event we encode mainly the type of action, the involved entities, the location and the timeframe in which an event takes place. Crucially, in GEST both explicit (e.g. actions) and implicit (e.g. existence) events are represented using the same rules. A complete example of GEST can be found in Figure 2.1. The causal relation between two events (i.e., E2 because E10) can be expressed as a single

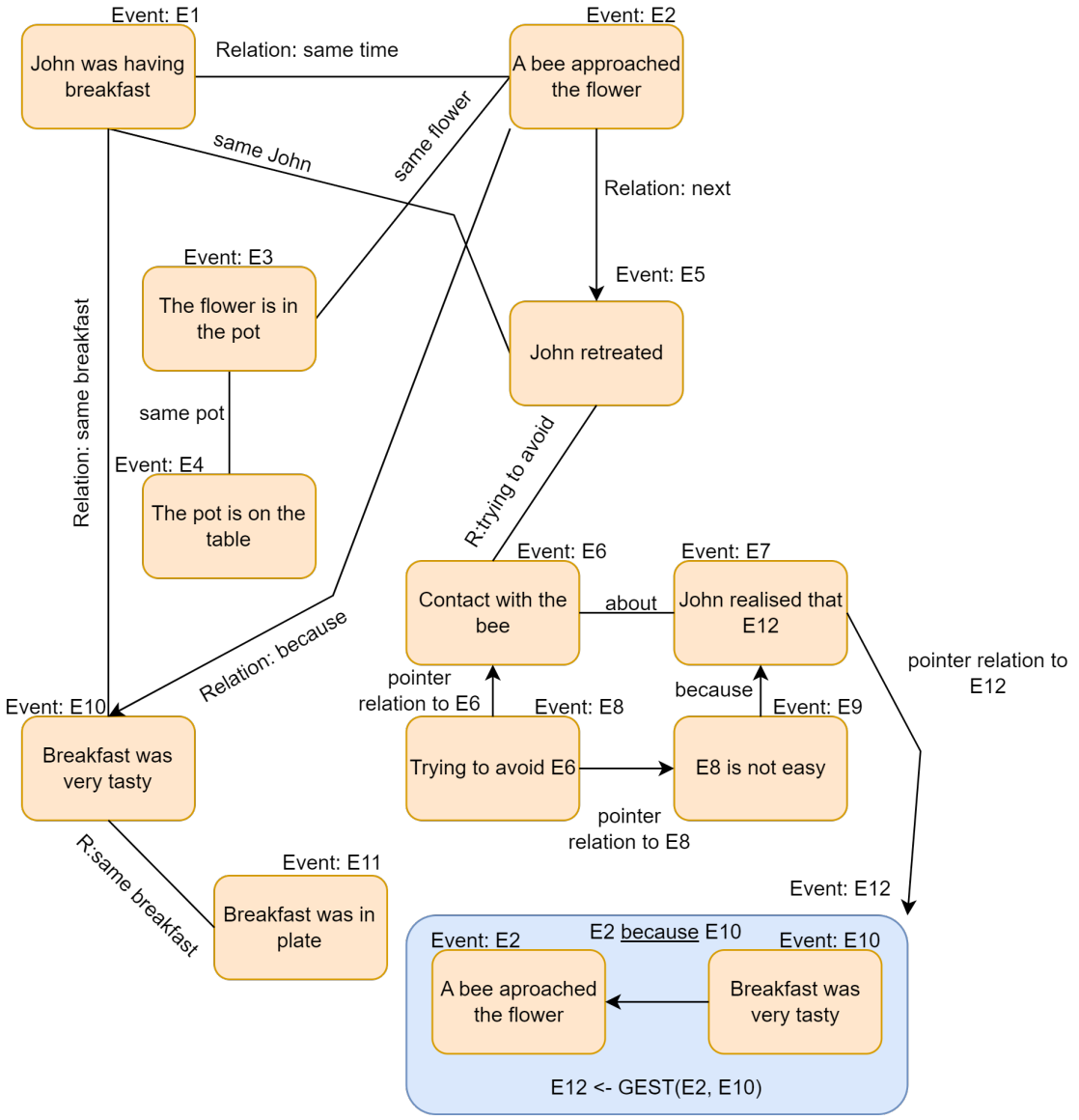

FIGURE 2.1: GEST graph explaining the following text: *"John was having breakfast when a bee approached the flower in the pot on the table. Then he pulled back trying to avoid contact with the bee but he realized that it was not an easy attempt because she actually came because of the tasty food on his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

event. In principle, any GEST could become an event into a higher-level GEST and vice-versa, any event could be expanded into a more detailed GEST.

Chapter 3

FROM TEXT TO GEST: GEST REPRESENTATION

Graphs of Events in Space and Time provide a common and meaningful representation for multiple modalities. In this chapter, we focus on building Graphs of Events in Space and Time from text representations. Translating texts into this novel representation spaces creates new opportunities. To prove GEST representational power and validate our approach we tackle the task of text similarity though a new point of view: instead of directly comparing texts, we bring texts in GEST space, and perform comparisons via graph matching metric in this novel space. We show that, for the task of detecting if two texts stem from the same video, our GEST based approach outperform classical text generation metrics and can also boost the performance of state of art, heavily trained metrics.

This chapter is based on the paper - *(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 Building ground truth GEST from text

Ground truth GEST from text is needed for training and evaluation. We note that building GEST representation from text is not a trivial task, and we aim to automate this process. Nevertheless, to obtain correct GEST from text human intervention is still needed. From each sentence, we want to extract information such as the type of actions, the entities involved, locations and the times of actions, as well as their relations. All this is information extracted by parsing the dependency tree (automatically extracted[1]) of each individual sentence using a set of handcrafted rules (followed if needed by human correction). Context (e.g. location inference) and event ordering is also injected into the graph to obtain the complete GEST of a story. While for the bAbI [5] corpus all entities (e.g. actors, objects) are unique for each story (e.g. a single actor with the name John in each story), in the case of Videos-to-Paragraphs [6] this is not always the case. Therefore, we have to manually intervene and set the proper references (build and link the proper number of nodes), as different entities are referred with the same name in the SVOs (e.g. ”man”, ”desk”). To find and accurately annotate these cases we manually go through each pair of SVOs and story and semantically check their validity.

3.2 Graph matching similarity metric

For comparing GESTs, we evaluate two graph matching methods, a classical approach, Spectral Matching (SM) [7] and a modern deep learning based approach, Neural Graph Matching (NGM) [8]. SM is a fast, robust and accurate method that uses the principal eigenvector of an affinity matrix, while NGM employs multiple neural networks that learn and transform the affinity matrix into the association graph, which is further embedded and used as input for a vertex classifier. For the following experiments,we consider all positive pairs from the test set of Videos-to-Paragraphs (a total of 67 in our case) and 174 negative pairs sampled randomly from the same test set.

[1] `https://spacy.io/models/en#en_core_web_lg` last accessed on 19^{th} of January 2023

For both SM and NGM algorithms, the affinity matrix is built using both node and edge level similarity functions that exploit pre-trained word embeddings. We use pre-trained GloVe [9] embeddings of size 300, to measure the similarity at each level (e.g. action, entities) for nodes. In order to compare two edges, we integrate node-level similarity (from the nodes that are connected to the particular edges) with the edge-level similarity (i.e. the similarity between the edges type). Essentially, two nodes are as similar as are their actions and entities, while the similarity of two edges is given by multiplying the edge type (e.g. next, meanwhile) similarity with the similarity of the corresponding nodes.

3.3 Results and Discussion

Results in Table 3.1 attest the power of our proposed representation: graph matching in the GEST space outperforms all classic text generation metrics (i.e. BLEU@4, METEOR and ROUGE) and even modern metrics based on pre-trained Transformers such as 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

TABLE 3.1: Results comparing GEST (ground truth) representation power with common text generation metrics applied on stories from Videos-to-Paragraphs test set. We show in **bold** the best value for each metric, and with <u>underline</u> the second best.

The initial test shows the representational power of GEST, but do not test the capability of this representation to be combined with a heavily trained one. We test this capability by showing that GEST can boost a state-of-the-art, strongly trained metric, even when we combine the two in the simplest, linear way. Starting from the original text of the story, we learn to automatically transform the story to GEST (via finetuning a

GPT3 model, text-curie-001[2]), and then obtain a GEST similarity score between stories by comparing, using graph matching, the corresponding generated GESTs. A second, BLEURT score between the stories is obtained as before. We then learn, on the training set, how to linearly combine the two scores, to best separate the texts of the same story vs. texts of different stories. We apply the same procedure to all classic metrics, in order to evaluate the benefit brought by GEST relative to other methods. In Table 3.2 we show the results of BLEURT (top), those of other metrics combined with BLEURT using the same linear regression approach (middle) and the results of GEST (bottom), using the two graph matching methods (SM and NGM). It is important to note that in combination with other metrics BLEURT does not always improve, but when combined with GEST it always improves and by the largest margin.

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>

TABLE 3.2: Results comparing the power of BLEURT coupled with common text generation metrics and GEST (learned), applied on stories from Videos-to-Paragraphs test set. Notations are the same as in Table 3.1.

Our experiments prove the power of GEST: its new space and associated graph matching metric can be effectively used, with minimal training cost, to boost the performance of existing state-of-the-art metrics. Even with very limited data, our experiments show that GEST is more than fitted for recreating the underlying story, within a space that allows for very reliable and human correlated comparisons.

[2] `https://platform.openai.com/docs/models/gpt-3` last accessed on 8^{th} of May 2023

Chapter 4

GENERATING VIDEOS FROM GEST

Artificial Intelligence makes great advances today and starts to bridge the gap between vision and language. However, we are still far from understanding, explaining and controlling explicitly the visual content from a linguistic perspective, because we still lack a common explainable representation between the two domains.

In this chapter we come to address this limitation and propose the Graph of Events in Space and Time (GEST), by which we can represent, create and explain, both visual and linguistic stories. We turn our attention to generating videos from GEST representations, in order to establish GEST as a common representation between vision and language. For that we will use a virtual environment, in which visual stories will be created from GEST. Note that once we can generate both videos and text from GEST, we will also be able to generate datasets of arbitrarily long and complex videos with

This chapter is based on the paper - *(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]*

full linguistic descriptions, which are currently very sparse but strongly needed in the literature.

4.1 From text to visual stories via GEST

To complete the connection between GEST and the visual world, we introduce the engine of visual stories. Based on the game GTA San Andreas with Multi Theft Auto (MTA)[1] interfacing the game's mechanics, we use the preexisting in-game locations, objects and animations and focus on events taking place in and around a house. The engine has full control within the virtual environment and can, therefore, take full advantage of the structured and explainable nature of GEST.

We can now generate datasets containing text-GEST-video triplets, which could be used e.g. to further advance the task of direct video to text translation. Now we are ready to build videos from GESTs. We note that this entire process is fully automated and requires no human intervention, thus enabling video generation at scale. The system takes a GEST as input and, based on it, generates multiple valid videos - note the one-to-many relation. This engine is used to automatically generate videos from GEST. Coupled with previously mentioned text-to-GEST module, we closed the loop and built a system that is capable of generation videos from text. Next, we generate a set of 25 complex videos of 2-3 minutes each, with up to 15 different activities, much larger than what is used in the current literature. Even if the set is small, it is very challenging so we use it to validate the our approach.

Next we present both human and automatic evaluations of our GEST-generated videos, compared to recent text-to-video models [10, 11]. It is important to note that both considered text-to-video models, CogVideo and Text2VideoZero are unable to generate videos that present more than one action, or in most case more than one actor. To ensure a fair comparison, we split the text into sentences, generate a video for each one, and concatenate them to get the final video.

[1] `https://multitheftauto.com/` last accessed on 25^{th} of July 2023

4.2 Evaluation

We invite human annotators to rate videos in terms of semantic content w.r.t input text, on a scale from 1 to 10 and pick the best video for each input text. In total, we collected 111 annotations. For each text, the users are presented with the original text the three generated videos and a battery of questions regarding the overall quality of each video and selecting the best one.

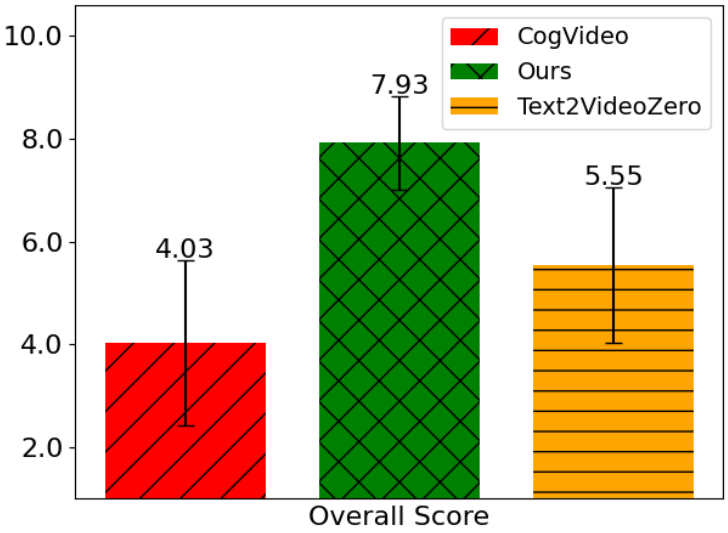

FIGURE 4.1: Overall scores (1-10) given by human evaluators.

Our generated videos are rated with an average score of 7.93, while videos generated using Text2VideoZero [11] and CogVideo [10] have an average score of 5.55 and 4.03 respectively, as displayed in Figure 4.1. In over 85% of cases, the human annotators picked our videos as overall best, with around 11% of cases in which Text2VideoZero is considered best.

We use VALOR [12] a recent video-to-text generation method, and measure how well the text generated back from the generated videos match the initial input texts. VALOR is trained and tested separately for each type of video generation method using 5-fold cross validation, from scratch, over 3 runs with results averaged (shown in Table 4.1). These experiments match the human evaluation, keeping the same ranking across methods and proving that GEST-generated videos can better maintain the semantic content of the original input text. This proves that an explicit and fully explainable vision-language

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

TABLE 4.1: Results on video-to-text task. We show in **bold** the best value for each metric.

model in the form of a graph of events in space and time, could also provide in practice a better way to explain and control semantic content - thus bringing a complementary value in the context of realistic (but not necessarily truthful) AI generation models.

Chapter 5

EXTRACTING GEST FROM VIDEOS

In previous chapters we have explored the power of GEST representation (through text-to-GEST task) and how GEST can be used to generate videos (i.e., GEST-to-video). In the following chapters we explore the missing pieces, namely how to build a GEST from a video (this chapter) and how to generate a rich description from a given GEST (the next chapter). An overview of our proposed approach is presented in Figure 5.1.

To construct an explicit representation of a video, we incorporate multiple tasks — primarily from the field of computer vision (action detection, semantic segmentation, depth estimation). For each frame in a given video, we first extract this information followed by a matching and aggregation step. The output of the action detector includes, for every action a bounding box of the person performing the action together with the name of the action and a confidence score. Starting from this bounding box, we aim to gather all the objects in the vicinity of the person, objects with which the actor could

This chapter is based on the paper - *(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]*

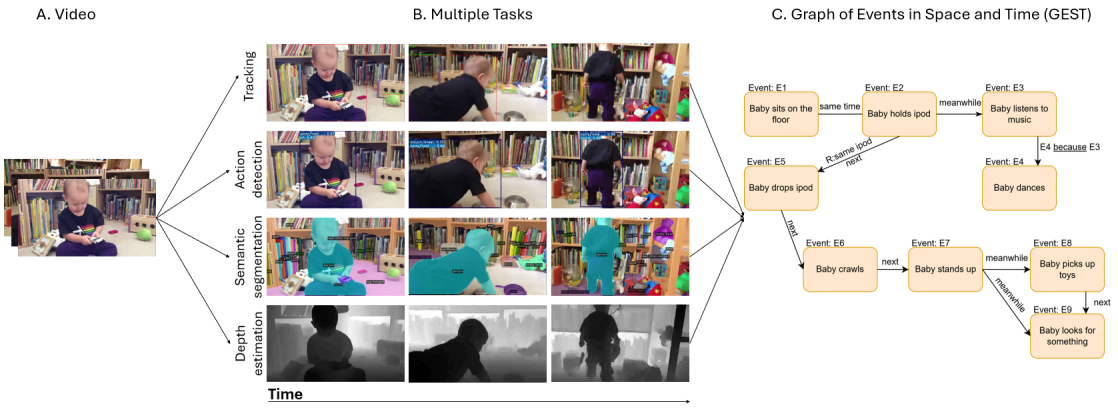

FIGURE 5.1: An overview of our approach. Starting from a raw video we perform object detection and tracking, action detection, semantic segmentation and depth estimation. We aggregate this information to build the corresponding Graph of Events in Space and Time.

interact. This list of objects is filtered based on depth difference between the person and the object to only. At this step we save for each action at each frame, information that includes the frame number, the person id (given at this point by the tracking model), the action name and confidence score as given the action detector, possibly involved objects and the bounding box of the person. Next, we apply frame-based action filtering, keeping only the top most high-confident actions, also implementing a voting mechanism with a window of 11 frames. The next step is to aggregate and process frame-level information into global video-level data. We solve short-term inconsistencies by unifying two person ids if they appear close in time (less than 10 frames) and they overlap enough (higher than 0.6 intersection over union), while long-term inconsistencies (such as when a person exits the frame and re-enters at a different stage and place) are solved via a semantic, appearance-based module that uses HSV histograms. After that, we aggregate actions that appear in consecutive frames (we allow a few missing frames) and save the starting and end frame, objects involved (union) and bounding boxes.

The last step in this entire pipeline in building spatio-temporal relationships between events: we build pairs of events and if they meet certain criteria we link them in space (measured euclidean distance of centroids) and time (next, same time or meanwhile).

Chapter 6

BUILDING RICH AND GROUNDED DESCRIPTIONS FROM GEST

As we previously tackled the video to GEST task, it is rather natural to discuss the task of describing the video in natural language. However, we stray away from the task of video captioning, as generally the language description in video captioning is very simple. In this chapter, we propose a common ground between vision and language based on GEST in an explainable and programmatic way, to connect learning-based vision and language state-of-the -rt models and provide a solution to the long standing problem of describing videos in rich natural language.

6.1 Method

Translating a GEST into a cohere, rich and natural language description is not a straightforward task with multiple possibilities. In this work we adopt a two-stage approach

Parts of this chapter are based on the paper - *(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]*

that harnesses the power of existing text-based LLMs to build natural descriptions. In the first step we convert the graph into a sound but maybe rough around the edges textual form, an initial description that we call proto-language. To obtain a more human-like description we use existing LLMs by feeding them with this proto-language and prompting with the goal of rewriting the text to make it sound more natural.

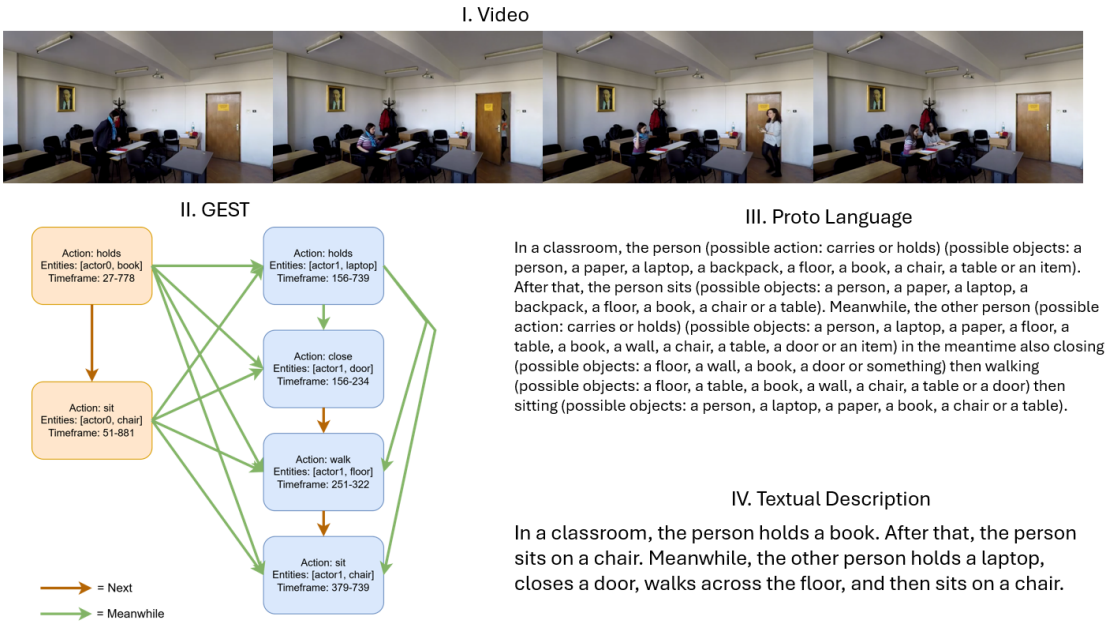

FIGURE 6.1: A complete example of our proposed pipeline. Starting from the video, we automatically build the associated GEST (see Chapter 5). From this graph, we build the proto-language that is then fed to an LLM that generates the final textual description.

Building the proto-language. The first step in this process involves a temporal sorting of the graph (by the start frame of each event; akin to a topological sort). Our approach aggregates chronologically sorted actions into higher-level groups of actions by actors. Each such group is then described in text, by describing each event using a simple grammar and taking into account the intra-group and inter-group spatial and temporal relations. When describing an event, we list all possible objects and let the LLM pick the objects that are most probable to appear in the given context, with the power to pick a new object that is not present in the list or not pick an object at all. Furthermore, we allow the LLM to change the name of an action or delete an action and its associated entities entirely if it does not fit the context. A complete example is presented in Figure 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)

TABLE 6.1: Absolute ordering (best is 1, worst is 5, followed by the absolute value of each metric) of the methods under different evaluations (i.e., text generation metrics, human evaluation and LLM-as-a-Jury) on the Videos-to-Paragraphs dataset.

Method	Average	VtP (489)	COIN (318)	WebVid (443)	VidOR (478)	VidVRD (164)
VidIL [14]	3.21	4.00	<u>2.82</u>	<u>2.93</u>	3.21	<u>3.11</u>
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	<u>3.16</u>	<u>1.96</u>	3.79	3.33	<u>3.16</u>	3.55
GEST + VidIL	**2.02**	**1.64**	**2.74**	**1.86**	**2.03**	**1.84**

TABLE 6.2: Average rank (best is 1, worst is 6) as selected by the LLM jury. VtP - Videos-to-Paragraphs. **Bold** marks the best result in each category, while <u>underline</u> marks the second best. The top performing method as evaluated using the LLM-as-a-Jury approach is again the combination between GEST and VidIL.

6.2 Results and Discussions

Results are presented in Table 6.1 and Table 6.2. On Videos-to-Paragraphs dataset, GEST clearly outperforms other methods, across all three evaluation directions. Enhancing our method with a more diverse set of actions and objects, or equivalently grounding the rich set of VidIL inputs with clear, concrete actions, both obtained simultaneously by combining GEST and VidIL, leads to better descriptions. Such descriptions are grounded, contain fewer hallucinations and better describe the source video.

In this chapter we introduce a novel method for accurately describing GEST in rich natural language. Combining this module with the Video-to-GEST module presented in the previous chapter we can tackle the story-like video description tasks.

Chapter 7

A LOOK INTO THE FUTURE - GEST AND (V)LLMS

In today's era on machine learning, where there is an abundance of data, both real and synthetic, not using heavily trained models would be a missed opportunity. Even though, as we shown in previous chapters, the quality of available data is not always optimal, state-of-the-art models still extract valuable insights from it and are capable of almost human-like performances. In this context, how does GEST align with the current landscape? Instead of competing against highly trained and already high performing models, we believe that GEST can serve as a complementary solution. And we already showed that this approach can work both when combining GEST with a state-of-the-art text generation metric (see Chapter 3) and when combining GEST with other methods for text-to-video task (see Chapter 6).

Parts of this chapter are based on the paper - *(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]*

Similarly, GEST could be used as a source of grounding in text-to-video and video-to-text pipelines. Text-to-image methods already started integrating extra information that grounds and controls the generated image. Lian et al. [19] make use of an off-the-shelf LLM to extract entities from the text prompt and build a grounded image layout. Next, a stable diffusion model guided by the generated layout is used to generate the final image. In the original implementation, both the layout generator and the image generator are general, off-the-shelf methods, with no further training or finetuning performed. Even though this is the inverse task, this approach is very similar to our proposed video-to-text method. Both share an intermediate representation (i.e., layout and proto-language) and both use off-the-shelf state-of-the-art methods and techniques. Zhang et al. [20] worked on adding control (e.g., pose, depth, segmentation) to text-to-image models while Ashual and Wolf [21] control the generated image via a scene-graph.

7.1 GEST and (V)LLMs

In this chapter, we are looking to investigate how can GEST be used in conjunction with state-of-the-art text-to-video methods, mainly Visual Large Language Models. Where and how could GEST be integrated? It is enough to simply add GEST in the prompt; should the models be trained or fine-tuned with GEST? What is the best way to integrate GEST, as a graph, as the proto-language or as the final description?

Multimodal Large Language Models (MLLM) build upon the success of (text-only) LLMs serving as a natural evolution that integrates multiple modalities. Visual Large Language Models (VLLMs) add vision capabilities to existing LLMs, allowing them to process, understand, and reason over visual input (i.e., images and videos). The development of VLLM followed a trajectory similar to that of LLMs, with Alayrac et al. [22] being the first to explore in-context learning for vision and language tasks, followed by the visual instruction fine-tuning paradigm [23] taking over the field.

To investigate if and how we can integrate GEST into existing VLLMs, we perform an initial set of experiments where, besides the video frames, we add either the proto-language (together with the ”assembly” instructions) or the final description. Results

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

TABLE 7.1: Videos-to-Paragraphs results when using the human annotated captions as ground truth. **Bold** marks the best result in each category, while underline marks the second best. B@4 stands for Bleu@4, M for Meteor, R for ROUGE-L, C for CIDEr, S for SPICE, BS for BERTScore and BT for BLEURT.

in Table 7.1 show that integrating such information directly into existing VLLMs is not straightforward. Adding the proto-language to the video input, barely increases the quality of the generated descriptions, only outperforming the baseline on two metrics (CIDEr and BERTScore). Using the final description yields even worse results, significantly underperforming compared to the baseline. We conjecture that this phenomenon stems from the training procedure of VLLMs, as they are not generally conditioned on multiple modalities for video captioning/description tasks. Therefore, we believe that the most effective methods for integrating GEST information into modern VLLMs is during the training phase, whether through pre-training, fine-tuning or in alignment phase.

Due to a combination of lack of models, data, hardware resources and know-how in training such large models at the start of this journey, we decide to start with a simpler, almost toy-like task. This greatly reduces the resources needed, both at hardware level and data level, allows us to better understand the process, the advantages and limitations of such models. Therefore, we pick the task of adapting VLLMs to Romanian language, as both a way to enable us a deeper understanding of these models and a way of giving back to the community.

Since any VLLM begins with a pre-trained visual encoder and a pre-trained LLM, the logical first step is to asses the effectiveness of existing models in Romanian. As we will show, particularly with earlier models (the only ones available at the beginning of this avenue of research), their performance in Romanian significantly lags behind they English proficiency across a wide range of tasks, with some models not even capable of answering a question in Romanian. This is expected as these models are mainly

English-focused with almost 90% of the data used for training being in English, while Romanian data accounted for only 0.03%.

For adapting existing LLMs and VLLMs to Romanian we resort to translating both training datasets and evaluation benchmarks from high-quality English sources. We also add natively Romanian downstream tasks and devise a novel benchmark that assesses the cultural (about Romania and Romanian culture) knowledge of LLMs.

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

TABLE 7.2: Comparison between RoLLMs and other LLMs on Romanian versions on academic benchmarks (abbreviations: HS - HellaSwag, Wino - Winogrande, G8k - GSM8k, TQA - TruthfulQA). **Bold** denotes the best in each category (average) and overall (each benchmark). We note the overall improvement of Romanian version across all families of models.

In Table 7.2 we present results on academic benchmarks for RoLLMs. We observe a consistent improvement of adapted RoLLMS compared to their original counterparts. For RoVLLMs, the results are presented in Table 7.3. Same as with RoLLMs, we observe a consistent improvements of Romanian VLLMs across the considered tasks.

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**

TABLE 7.3: VLLM performance on Romanian benchmarks. *Italic* marks model built from "scratch": adapter randomly initialized, base LLM as is. The baselines (non-italic) have already been aligned and finetuned on English data. Note that models marked with *italic* in the lower section are finetuned versions of models presented in the middle section (aligned models). **Bold** marks best result for each benchmark.

In this chapter we introduced the first open-source LLMs and VLLMs specialized for Romanian. Our evaluations show promising results, outperforming existing solution across multiple benchmarks. For both LLMs and VLLMs, we present a general training and evaluation recipe, including resources, recipe that we expect to work on other architecture as well. Same as with RoLLMs, we present a general training and evaluation recipe, including resources, recipe that we expect to work on other, larger and more powerful architectures.

Future directions are shared with RoLLMs, as we mainly use the same approach: increasing the number and quality of datasets, by validating translation and collecting human instructions and preference data. Furthermore, for RoVLLMS, at this stage we did not perform human preference alignment on the current models, a stage we showed was critical for LLM performance.

Chapter 8

CONCLUSION AND FUTURE WORK

This thesis has introduced a novel framework and representation for bridging together vision and language. We have provided both theoretical foundations and strong practical methods and results, that can be extended to other modalities. Furthermore we did not tackle all tasks at the intersection and vision and language (e.g., visual question answering), and we are sure that more interesting applications of this framework exist and will be studied (e.g., reasoning in GEST space). Throughout our research, we developed novel approaches, building valuable resources along the way. It is important to note that this works only represents the first steps in defining and utilizing GEST at its full potential and numerous avenues for research are still open.

Below, we provide a summary of the contributions this thesis brings:

Representation. We introduce a novel, graph-based representation in the form of Graph of Events in Space and Time. We create GEST with the main goal of building a universal representation of sequences of events, of stories. We consider this representation close to how stories are represented and used in the human brain. We provide the intuition behind GEST, a formal definitions and examples of its capability and universality and show how can be we build GESTs from text.

Text-to-Video Generation. Having already generated GESTs from text, the next task in line was GEST-to-video generation. Using existing game engine resource (MTA), we devised a procedure to build a visual story from a GEST. The explicit nature of GEST combined with the algorithmic approach guarantees the accuracy of the generated video, irrespective of length and complexity.

Rich video description. We have also explored the inverse task, from video to text. Exploiting the rich nature of GEST, we focused on the story-like video description task, as opposed to the simpler task of video captioning. Harnessing a multi-task approach we extract from each video, a grounded, rich and complete Graph of Events in Space and Time (Chapter 5). The second stage in the video-to-text pipeline, namely GEST-to-text, is tackled using a combination of graph sorting (in space and time), grammar-based rules and LLMs to accurately depict in text the events that are present in GEST.

Combining GEST with VLLMs. Inspired by previous experiments, where we show that GEST is complementary to state-of-the-art solutions for text generation metrics and video descriptions we start to investigate if and how GEST can be combined with largely unexplainable but very powerful Visual Large Language Models. Initial experiments showed that integrating an extra modality in the form of GEST (even if in textual form) is not really straightforward, so we started on a mission to better understand how to train LLMs and how to best add extra modalities.

8.1 Future Research Agenda

We consider that our journey so far represents only the beginning of GEST and its applications. This thesis just laid the foundations and identified part of the challenges towards an explainable bridge between vision and language. We hope that this inspires and paves the way for other researchers to overcome the limitations of our work and open up new horizons for vision and language tasks.

Next, we outline some research directions that represent natural extensions of our work, that were not yet investigated due to time and resource limitations.

- **Increase the number of tasks.** While we proposed solutions for two vision and language tasks, video description and text-to-video generation, we did not discuss how GEST can be used to solve other tasks and the intersection of vision and language, such as visual question answer (VQA) or visual entailment (VE) [24]. In VQA the goal is to answer a natural language question about an image or a video. For this task GEST could be instrumental, as a possible pipeline could include a multi-step process that converts the video into a GEST and the natural question in a query, that can be used to interrogate the graph. In general, in this thesis we used the characteristic of GEST for grounding, we did not particularly focus on applications involving reasoning over GEST or even altering the graph (i.e., for adversarial examples generation).

- **Integrating GEST in VLLMs and Diffusion Models.** As already shown in this thesis, information encoded in GEST can complement existing state-of-the-art solutions. Adding grounding via GEST in either VLLMs or Diffusion models could significantly increase the complexity and accuracy of the generated texts or videos. We already built valuable resources, including tuples of video-GEST-ranked list of textual descriptions, that can be directly used (e.g., as human preference dataset for video description).

- **Improving and expanding existing resources.** Connecting to the previous direction, we have already generated videos starting from text. Using this data and applying existing techniques such as semantic segmentation to extract person bounding boxes, we could generate layouts for a visual generation model (e.g., Diffusion Models). Further expanding the generated resources, both in number and quality, can be used to further improve existing methods. Finally, each of the methods developed, implemented and evaluated in this thesis holds potential for further improvements, particularly given the impressive advancements in LLMs and VLLMs.

8.2 Closing remarks

In this era where artificial intelligence has become ubiquitous, AI assistants now operate seamlessly on hand-held devices, capable of integrating and generating video, speech and text. Nowadays, AI is being applied across virtually every domain – medicine, law, education, sports, politics and more. While these models demonstrate remarkable capabilities in a vast array of tasks and will certainly continue to evolve, we believe we have a responsibility to study and invest in AI safety and explainability.

Beyond achieving and even surpassing human performance of benchmarks, an impressive feat in itself, it is imperative that we also strive to interpret and understand the decision making processes of such AI models. And explainability comes in a variety of ways, including approaches such as exposing reasoning tokens [25], input perturbation [26], or gradient-based techniques [27]. Moreover, in a worlds where AI is more than likely used in high-stake domains including military applications, AI safety becomes paramount.

We hope this work contributes to and inspires the research community, guiding the development of fair, safe, unbiased, explainable and trustworthy AI systems.

Bibliography

[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.