Explaining Vision and Language through Graphs of Events in Space and Time

Mihai Masala[1,2]   Nicolae Cudlenco[1]   Traian Rebedea[2]   Marius Leordeanu[1,2] *

[1] Institute of Mathematics of the Romanian Academy   [2]University Politehnica of Bucharest

Abstract

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 work 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 provide a theoretical justification of our model and an experimental validation, which proves that GEST can bring a solid complementary value along powerful deep learning models. In particular, GEST can help improve at the content-level the generation of videos from text, by being easily incorporated into our novel video generation engine. Additionally, by using efficient graph matching techniques, the GEST graphs can also improve the comparisons between texts at the semantic level.

1. Introduction

There is a considerable amount of research at the intersection of vision and language, such as image and video generation [22, 40, 15, 3, 33, 24, 29], captioning [11, 39, 30] or visual question answering [2, 18, 38]. However, we still lack an explainable model that can fully relate, constrain and control the connection between vision and language at the level of meaning and content. This limitation, which affects not only text-to-image/video models, but also Large Language Models [36], seriously impedes our way towards trustworthy and safe AI. We mention that, even in this work, we found state of the art text-to-video transformer models generating almost adult-only content for a simple, plain text such as: *A woman goes to the bedroom.*

In this context, we introduce GEST, the Graph of Events in Space and Time, which provides an explicit spatio-
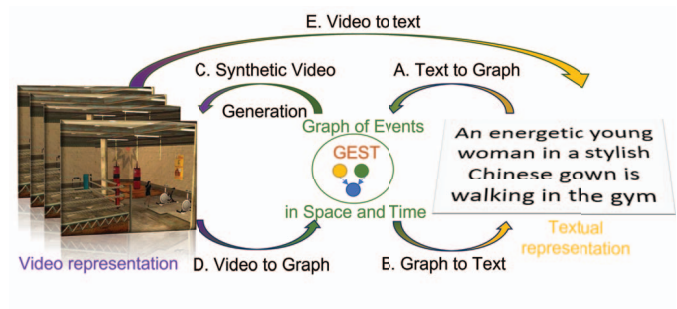

Figure 1. Functional overview of the proposed framework. GEST represents the central component, allowing for the preservation of the semantic content in an explainable form, as well as a seamless transition between different domains.

temporal representation of stories as they appear in both videos and texts and can immediately relate, in an explainable way, the two domains. GEST provides a meaningful representation space, in which similarities between videos and texts can be computed at the level of semantic content. GEST can also be used in the context of our specially designed video generation engine (Sec. 3) to produce videos that are rated higher in terms of content, both by human and automatic evaluations, than their video counterparts generated by state of the art text-to-video models (Sec. 4). Also, GEST graphs can be used for comparing the meaning of texts and improve over classic text similarity metrics or in combination with heavily trained state-of-the-art deep learning metrics (Sec. 2.1). Graphs have been used to represent content in videos [26, 25, 7, 34, 32, 9] or texts [17, 35, 19, 10, 4], but not both as is the case for GEST.

Main novel aspects of GEST are: **1)** Nodes are events, which could represent (Sec. 2) physical objects, simple actions or even complex activities and stories. **2)** Edges can represent any type of relation (temporal, spatial, semantic, as defined by any verb) between two events defined as nodes. **3)** Any GEST graph can always collapse into a node event, at a higher level of abstraction. Also, any event node can always be expanded into a GEST graph, from a lower

---

*Primary contact: Marius Leordeanu at leordeanu@gmail.com

level of abstraction. This is an essential property that allows GEST to have multiple layers of depth (see Fig. 2).

Another practical **contribution** of our work, is our novel video generation engine (Sec. 3), based on GEST, which can produce long and complex videos that preserve well semantic content, as validated by human and automatic evaluations. We will make the engine code and the videos generated for our experiments publicly available.

2. GEST Model

The basic elements of GEST are the nodes, which represent events and the edges, which represent the way in which events interact.

GEST nodes: represent events that could go from simple actions (e.g. opening a door) to complex, high-level events (e.g. a political revolution), in terms of spatio-temporal extent, scale and semantics. They are usually confined to a specific time period (e.g. a precise millisecond or whole year) and space region (e.g. a certain room or entire country). Events could exist at different levels of semantics, ranging from simple physical contact (e.g. "I touch the door handle") to profoundly semantic ones (e.g. "the government has fallen" or "John fell in love with physics"). Even physical objects are also events (e.g. John's car is represented by the event "John's car exists"). Generally, any space-time entity could be a GEST event.

GEST edges: relate two events and can define any kind of interaction between them, from simple temporal ordering (e.g. "the door opened" after "I touched the door handle") to highly semantic (e.g. "the revolution" caused "the fall of the government", or "Einstein's discovery" inspired "John to fall in love with physics"). Generally, any verb that relates two events or entities could be a GEST edge.

From a graph to a node and vice-versa: A GEST graph essentially represents a story in space and time, which could be arbitrarily complex or simple. 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.

Going from a $GEST$ at a lower level to an event E at a higher level ($E \Leftarrow GEST$) reminds of how the attention mechanism is applied in Graph Neural Networks and Transformers [27]: the GEST graph acts as a function that aggregates information from nodes (events) E_i's at level k
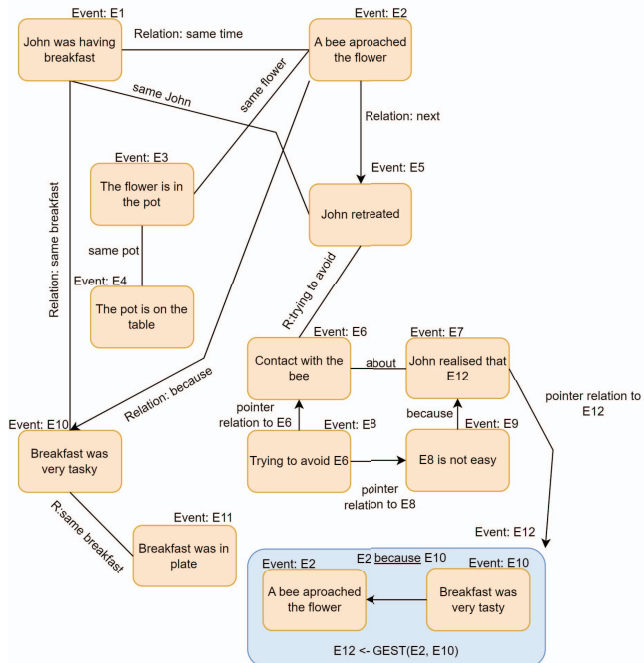

Figure 2. GEST graph explaining the following text: *"John was having breakfast when a bee approached the flower in the pot on the table. Then he pulled back trying to avoid contact with the bee but he realized that it was not an easy attempt because she actually came because of the tasty food on his plate"*.

and builds a higher level GEST representation, which further becomes an event at the next level $k + 1$:

$$E_i^{(k+1)} \Leftarrow GEST(E_1^k, E_2^k, ..., E_n^k)$$

In Fig. 2 we present our GEST representation, as it applies to a specific text. In each event node, E_i, we encode an $action$, a list of $entities$ that are involved in the action, its $location$ and $timeframe$ and any additional $properties$. Note that an event can contain references (pointers) to other events, which define relations of type "same X" (e.g. "same breakfast"). We also exemplify how the GEST of two connected events can collapse into a single event.

2.1. GEST for Textual Content Comparison

Next we verify experimentally that the GEST model can capture the semantics of language by applying it to the task of text to text comparisons, in the context of video to text translation. We use the Videos-to-Paragraphs dataset [6] that has multiple text descriptions for the same video. Starting from the given texts, we build ground truth GEST representations for the entire dataset as follows: we use a rule-based method to obtain initial GESTs from texts, represented in a specific string format that captures information in the nodes as well as their relationships. Next we check, correct and refine the automatically generated GESTs by human annotation. Note that we also tested with GhatGPT,

Method	Corr(%)	Acc(%)	F	AUC(%)
BLEU@4	24.45	75.52	0.28	52.65
METEOR	58.48	84.23	1.12	73.90
ROUGE	51.11	83.40	0.72	68.92
SPICE	59.42	84.65	1.04	74.43
BERTScore	57.39	85.89	1.07	**77.93**
GEST-SM	**61.70**	84.65	**1.20**	75.47
GEST-NGM	60.93	**86.31**	0.98	76.75

Table 1. Comparing GEST representation power (coupled with graph matching similarity functions SM or NGM) and well-known text-to-text similarity methods (applied on texts from Videos-to-Paragraphs test set, on the task of separating texts describing the same video vs. texts from different videos). Corr - correlation, Acc - Accuracy, F - Fisher score and AUC - area under the precision-recall curve. Best values are in **bold**, second best underlined.

which was able to produce mostly valid GESTs by learning from a few human examples.

We seek to find how useful is GEST in deciding if two texts stem from the same video or not. Basically, instead of comparing texts, we move the comparison in the GEST space in which we define a similarity function using graph matching. In We use as graph matching methods the classic Spectral Matching (SM) [14] and the recent Neural Graph Matching (NGM) [31]. For both algorithms, the affinity matrix is build using node and edge level similarity functions based on pre-trained GloVe [21] word embeddings. Two nodes are as similar as are their components (e.g. action, entities), while edge-level similarity uses the relation type defined by the edge (e.g. causality, temporal ordering, etc.) along with the similarity of the nodes they connect.

In Tab. 1 we present comparisons of GEST+graph matching similarity vs. other well-known text similarity metrics, which demonstrate that GEST is capable to capture semantic content. In Tab. 2 we investigate whether graph matching in GEST space can be combined with state-of-the-art highly trained text similarity metrics such as BLEURT [23]. We combine each pair of similarity metrics (BLEURT + X) in linear way, to ensure that if a performance gain exists, it is less likely to be due to the combination method and more due to the additional metric. In this setting GEST graphs are learned by finetuning a GPT-3 model (*text-curie-001*), with raw text as input and ground truth GEST as output, on the Videos-to-Paragraphs train set. Note that the combination of BLEURT with graph matching in the GEST space consistently increases the performance over BLEURT (which is not always the case for other metrics) and by the largest margin.

3. GEST Video Generation Engine

To complete the connection between GEST and the visual world, we introduce the engine of visual stories. Based

Method	Corr(%)	Acc(%)	F	AUC(%)
BLEURT	70.93	90.04	2.03	88.02
+BLEU@4	70.93	90.04	2.03	88.04
+METEOR	71.20	89.63	2.07	87.62
+ROUGE	70.76	90.04	2.00	87.71
+SPICE	71.94	88.80	2.09	87.71
+BERTScore	71.11	89.63	2.01	87.25
+GEST-SM	**72.89**	**90.87**	**2.21**	**89.80**
+GEST-NGM	71.91	90.46	2.05	88.58

Table 2. Results comparing the power of BLEURT coupled with well-known text similarity metrics and GEST, applied on stories from Videos-to-Paragraphs test set. Text metrics are computed on the ground truth stories, while GESTs are generated with a transformer learned on the training set. Same notations as in Tab. 1.

on the game GTA San Andreas with Multi Theft Auto (MTA)[1] interfacing the game's mechanics, we use the pre-existing 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. It is capable of choosing a setting in a virtual environment, with locations, actions and entities that match the events described within the GEST and orchestrate the complex interactions during the simulation, thus emulating an entire world (Figure 3).

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. We couple this with the system that generates GEST starting from a text, closing the loop and building a system that transforms a text into a GEST, then a GEST into a video. 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 to validate the quality of the generated videos. Results of this evaluation are presented in the following section.

Metric	Ours	CogVideo	Text2VideoZero
Bleu@4[20]	9.84	8.16	**10.02**
Meteor[5]	**14.16**	13.48	13.96
ROUGE[16]	**35.40**	32.72	34.87
SPICE[1]	**20.04**	19.54	19.43
CIDEr[28]	**34.12**	33.16	33.65
BERTScore[37]	**19.37**	13.09	15.02
BLEURT[23]	**39.44**	37.55	38.40

Table 3. Results on video-to-text task. We show in **bold** the best value for each metric.

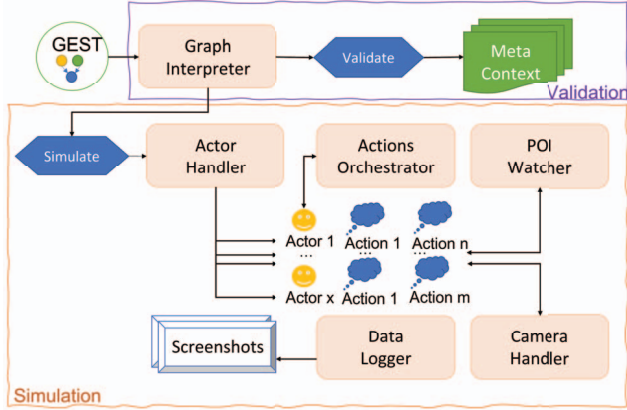[1] https://multitheftauto.com/, accessed on 25 July 2023

Figure 3. The system architecture of the engine. Upper part - meta context validation. Lower part - simulation.

4. Vision-Language Experiments with GEST

Next we present both human and automatic evaluations of our GEST-generated videos, compared to recent text-to-video models [12, 13]. 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. We collected a total of 111 annotations, from 6 independent annotators. In Fig 5 we show the overall scores given by human evaluators for each method. In 87.39% of cases our GEST-generated video was picked as best, with only 11.71% for Text2VideoZero and 0.90% for CogVideo.

For the automatic evaluation of the generated videos, we use a state-of-the-art video-to-text generation method, VALOR [8], 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 Tab. 3). 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 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.

The reason why current deep learning models are not strong is that we generate long and complex videos. Their main weakness resides in their inability to integrate long and complex context, both in video and text generation.

5. Conclusions

We propose an explainable representation that connects language and vision (see Fig 1), which explicitly captures semantic content as a graph of events in space and time
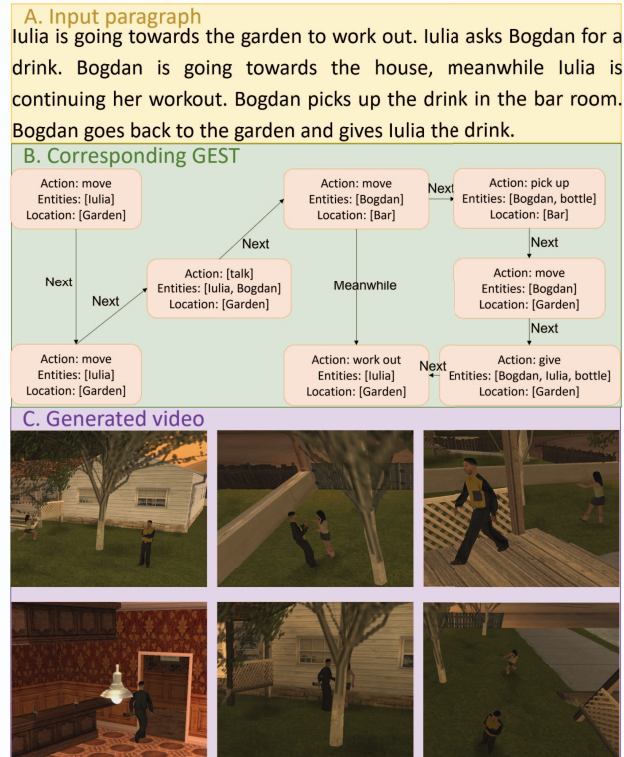

Figure 4. Example of input text (A), generated GEST from text (B) and automatically generated video from GEST (C).
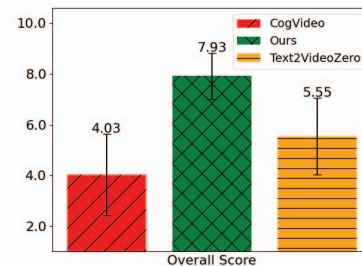

Figure 5. Overall scores (1-10) given by human evaluators.

(GEST). We prove that GEST is capable of capturing meaning from text and contribute to the design of powerful text-to-text comparison metrics when combined with graph matching. More importantly, GEST can be also used to generate videos from text that better preserve the semantic content (as evaluated by humans and automatic procedures), than deep learning methods for which there is no explicit way of explaining and controlling content. In future work we plan to explore ways to better integrate the power of deep learning into the explainable structure of GEST, for further developing a robust and trustworthy bridge between vision and language.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[3] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2, 2019. 1

[4] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186, 2013. 1

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. 3

[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. 2

[7] William Brendel and Sinisa Todorovic. Learning spatiotemporal graphs of human activities. In *2011 International Conference on Computer Vision*, pages 778–785. IEEE, 2011. 1

[8] 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. 4

[9] Anoop Cherian, Chiori Hori, Tim K Marks, and Jonathan Le Roux. (2.5+ 1) d spatio-temporal scene graphs for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 444–453, 2022. 1

[10] Janara Christensen, Stephen Soderland, Oren Etzioni, et al. Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1173, 2013. 1

[11] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017. 1

[12] 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. 4

[13] 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. 4

[14] 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. 3

[15] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1

[16] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 3

[17] Dekang Lin. A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering*, 4(2):97–114, 1998. 1

[18] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016. 1

[19] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988. 1

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 3

[21] 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. 3

[22] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 1

[23] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. 3

[24] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[25] Dinesh Singh and C Krishna Mohan. Graph formulation of video activities for abnormal activity recognition. *Pattern Recognition*, 65:265–272, 2017. 1

[26] Muralikrishna Sridhar, Anthony G Cohn, and David C Hogg. Relational graph mining for learning events from video. *STAIRS 2010*, pages 315–327, 2010. 1

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[28] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 3

[29] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 1

[30] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018. 1

[31] 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. 3

[32] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018. 1

[33] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pretraining for neural visual world creation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022. 1

[34] Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, and Abhinav Gupta. Temporal dynamic graph lstm for action-driven video object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1801–1810, 2017. 1

[35] Luke S Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*, 2012. 1

[36] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023. 1

[37] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 3

[38] Huasong Zhong, Jingyuan Chen, Chen Shen, Hanwang Zhang, Jianqiang Huang, and Xian-Sheng Hua. Self-adaptive neural module transformer for visual question answering. *IEEE Transactions on Multimedia*, 23:1264–1273, 2020. 1

[39] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 1

[40] Xingran Zhou, Siyu Huang, Bin Li, Yingming Li, Jiachen Li, and Zhongfei Zhang. Text guided person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3663–3672, 2019. 1