Compute to Tell the Tale: Goal-Driven Narrative Generation

Yongkang Wong
Shaojing Fan
National University of Singapore

Yangyang Guo
Ziwei Xu
National University of Singapore

Karen Stephen
NEC Corporation

Rishabh Sheoran
National University of Singapore

Anusha Bhamidipati
Vivek Barsopia
Jianquan Liu
NEC Corporation

Mohan Kankanhalli
National University of Singapore

ABSTRACT

Man is by nature a social animal. One important facet of human evolution is through narrative imagination, be it fictional or factual, and to tell the tale to other individuals. The factual narrative, such as news, journalism, field report, etc., is based on real-world events and often requires extensive human efforts to create. In the era of big data where video capture devices are commonly available everywhere, a massive amount of raw videos (including life-logging, dashcam or surveillance footage) are generated daily. As a result, it is rather impossible for humans to digest and analyze these video data. This paper reviews the problem of computational narrative generation where a goal-driven narrative (in the form of text with or without video) is generated from a single or multiple long videos. Importantly, the narrative generation problem makes itself distinguished from the existing literature by its focus on a comprehensive understanding of user goal, narrative structure and open-domain input. We tentatively outline a general narrative generation framework and discuss the potential research problems and challenges in this direction. Informed by the real-world impact of narrative generation, we then illustrate several practical use cases in Video Logging as a Service platform which enables users to get more out of the data through a goal-driven intelligent storytelling AI agent.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **General and reference** → *Surveys and overviews*; • **Applied computing**;

KEYWORDS

Computational Narrative Generation, Video Analytic

ACM Reference Format:
Yongkang Wong, Shaojing Fan, Yangyang Guo, Ziwei Xu, Karen Stephen, Rishabh Sheoran, Anusha Bhamidipati, Vivek Barsopia, Jianquan Liu, and Mohan Kankanhalli. 2022. Compute to Tell the Tale: Goal-Driven Narrative Generation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3503161.3549202

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3549202

Figure 1: A brief overview of narration in human history. Left: Cave paintings were found on cave walls that date to over 10,000 years ago. Middle: Oracle bone script was the ancestor of modern Chinese characters dated at 1300 to 1100 BCE. Right: Theatrical performance consists of dialogue between actors that is enriched with music and choreography.

1 INTRODUCTION

The amount of video data generated everyday has seen a rapid increase in recent years — largely due to the availability of affordable and portable video capture devices, cloud-based storage and streaming services (*e.g.*, iCloud, YouTube, TikTok, etc.). In modern times, the typical ways that people consume digital data are through content retrieval, such as via search service [28] or recommender systems [1, 32, 54, 108], or by interacting with intelligent AI agents via natural language [3, 15, 29, 61, 84]. Notably, the raw data need to be first curated by human experts or processed by advanced AI models, which is often challenging due to scalability, creativity, and content diversity. Mirrored to one of the most ubiquitous media, *i.e.*, videos (either long or short), it is rare for this kind of data to be released without any additional editing. As such, building a computational system to generate a high quality summary from raw videos has become increasingly demanding thus far.

In the history of human civilization, narrative is regarded as the basic vehicle of human knowledge transfer [62] and it is important for its social, cognitive, and cultural roles in the lives of storytellers and receivers [13]. It serves as an effective tool for communication and building relationships, as well as transforming the ways that humans understand and see the world [5]. Narration — the art of storytelling — has evolved from cave drawing to linguistic narrative text (see Fig. 1), and has grown into theatrical performance through the ages. In present day, films, plays, novels, biography, newsreels and chronicles are all narratives as they all connect series of events in a causal manner [44]. However, the creation of a good narrative is non-trivial, where a compelling narrative requires an interesting story structure, coherence, informativeness, temporality, causality, and attracting receivers' interest [13]. In other words, narrative generation is a complex task that takes humans years of training to master. In this paper, we discuss a fundamental question — can an artificial computational system generate narratives?

This study focuses on a particular use case of narrative generation, namely **goal-driven narrative generation with long videos**, where a **factual narrative**, *i.e.*, stories based on actual events, is generated. The proposed task takes a user's query and multiple long videos (ranging from an hour to a few days) as input, and then generates a coherent and succinct multimodal narrative. We note that the research community has made good preliminary progress through image/video captioning [39, 95], image paragraph generation [99], story ending generation [27, 36] and video storytelling [48, 103]. However, a fully computational narrative generation system is still a challenging and ambitious goal. Not only it requires robust perceptual capability in both seen and unseen scenes, it also demands a comprehensive understanding of user goal, narrative structure and open-domain input. Most importantly, **there exist no single objective story** and what is deemed to be interesting varies across users, topics, and other factors. To help tackle this daunting problem, we propose a research agenda through a general processing framework and discuss the potential research challenges. In addition, we review existing effort from both social science and computer science perspectives. To summarize, we make the following contributions:

- We introduce a novel goal-driven computational narrative generation task for long videos. A general framework is outlined to encourage domain and style agnostic approach. The anticipated research challenges are discussed.
- We present a targeted survey on the social science studies on narratives and bridge the literature to the proposed task. In addition, we compare the proposed task with existing literature and outline the key differentiating factors.
- We delineate several conceptual applications of user-centered narrative generation in Video Logging as a Service (VLaaS) platform, such as smart warehouse, claim report generation, and unmanned region monitoring.

In the remaining sections, Section 2 positions narrative with findings from social science studies. Section 3 covers the computational narrative generation task where the comparison with related tasks (*cf.* Section 4) and related research problems (*cf.* Section 5) are also detailed. We complement these contributions by discussing how narrative generation application can impact our society in Section 6.

2 SOCIAL SCIENCE ASPECT OF NARRATIVE

"Narrative imagining — story — is the fundamental instrument of thought. [...] It is a literary capacity indispensable to human cognition generally".

— by Mark Turner

In social sciences, narratives have been a long studied topic since the late 1980s. Narrative is regarded as the basic vehicle of human knowledge [62]. In particular, a narrative is defined as a semiotic representation of a series of events connected in a temporal and causal way. Films, plays, novels, newsreels and chronicles are all narratives in this widest sense [44]. In this study, we will use the word in a more restricted sense, meaning a linguistic narrative text or a visual video, or the representation of a series of events by means of textual or visual modalities. While in literature most of the events are fictional, our narratives will describe nonfictional (*i.e.*, factual) events based on the information from the videos.

A linguistic narrative structure has multiple modalities, the two most important modalities are **temporal** and **spatial** [62]. The temporal modality refers to the representation of events in a time sequence in a narrative. From the temporal perspective, a narrative can be seen as a sequence of sentences that flow along in time one after the other. But the sentences in a narrative contract with each other more than merely temporal relations, instead, they can be conceptually and thematically related to each other, for example, with causal relations. The **causal** modality indicates the causal connections between the narrated events. Cohn [11] has stated the definition of narrative from the causal angle: "a series of statements that deal with a causally related sequence of events that concern human (or human-like) beings".

In this study, we target generating narratives with both temporal and causal modalities from videos. For example, the following is a series of video scenes with only temporal modality: "A man with a bag passes by a bench in a park. A bag is on the bench. A police car comes". Alone, this is a non-narrative sequence. But an intelligent algorithm can postulate connections that would weave these scenes together into a narrative: "A man abandons a bag on the bench in a park. Someone nearby spots the abandoned bag and calls the police. The police drives to the park to take a look". In this example, causal ties are necessary to make the narrative a complete story.

Besides the aforementioned modalities, researchers argue that narrative also has a spatial structure [6, 22, 72]. The sentences in a narrative can be simultaneously related to sentences both preceding and following them in complex ways, gives narratives a spatial configuration. This lays the foundations for narrative diagrams semiotic, which can be viewed as complex networks of relations that exist simultaneously in a mental, multidimensional space.

There are many factors that influence the narrative structure, among which the most prominent is the narrator's goal. As the saying goes, "what we say is not as important as how we say it", research has shown that goal, or motivation, has strong impact on the narrative construction [23]. The flexibility and complexity associated with narrative construction enables us to present the information in the structure of our choice. The versatility of narrative structure allows us to convey a variety of detailed information with specific targets and from multiple perspectives. For example, a structural consideration of narrative construction in our algorithm may cause a person to focus on certain story elements while ignoring others.

Indeed, it is well-established that narratives can have effects on readers' real-world beliefs and attitudes [4]. This phenomenon has been termed **narrative persuasion**. In communication research, such effects have long been investigated in various disciplines such as health communication [25] and entertainment-education [56]. In strategic political communication and policymaking, the use of narratives and storytelling are also recognized as effective communication strategies [57]. Narrative persuasion is also linked to the evoking of various emotions [34]. The above research provides insights for the design of computational narrative generation, such as the adjustment of narrative structure based on different motivations and target impacts.

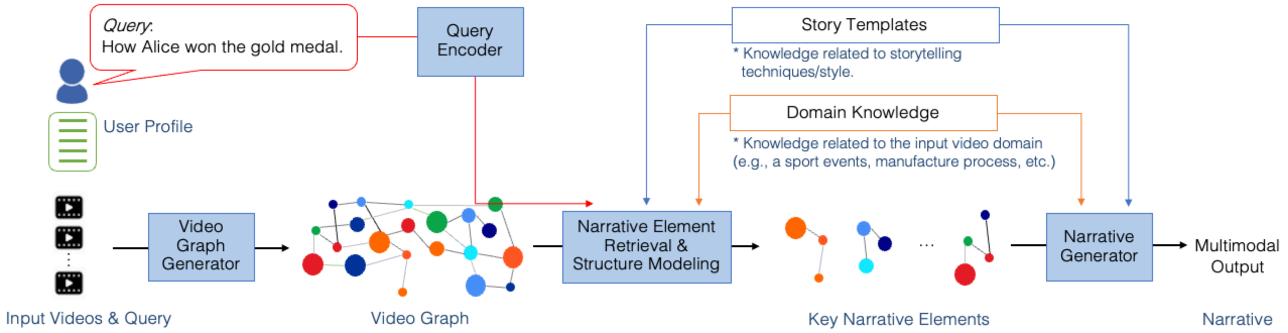

Query:
How Alice won the gold medal.
User Profile
Query Encoder
Story Templates
* Knowledge related to storytelling techniques/style.
Domain Knowledge
* Knowledge related to the input video domain (e.g., a sport events, manufacture process, etc.)
Video Graph Generator
Narrative Element Retrieval & Structure Modeling
Narrative Generator
Multimodal Output
Input Videos & Query
Video Graph
Key Narrative Elements
Narrative

Figure 2: An overview of a general narrative generation framework with input video and user’s query. Firstly, the Video Graph Generator processes the input video into a comprehensive semantic video graph. Secondly, the Narrative Element Retrieval determine what make an interesting story based on the given query, story templates, and domain knowledge. Finally, the Narrative Generator output a textual and/or video story corresponding to the user’s query. The story templates and domain knowledge are explicit knowledge that determine the narrative elements.

3 NARRATIVE GENERATION FRAMEWORK

3.1 Problem Definition

The aim of narrative generation is to generate a coherent and succinct story from a single or multiple long videos. Specifically, we consider both raw videos and non-compressed narrated videos, *e.g.*, movie, TV series, YouTube video, etc. In this paper, we consider a general narrative generation (NG) framework for long videos employing the conventional deep learning terminology. Given a multimodal query Q concerning a set of long videos $\mathcal{V} = \{V_1, V_2, \dots, V_K\}$ where K is the number of videos, the goal of NG is to generate a narrative — story — in linguistic narrative text $S_{\texttt{nar}}$ with an optional narrative video $V_{\texttt{nar}}$ to complement it.

An overview of the general NG framework is illustrated in Fig. 2. Briefly, the pre-processing stage converts the input $\mathcal{V}$ into a structured semantic video graph and retrieves the related domain knowledge $\mathcal{K}_{\texttt{dom}}$ based on Q. This stage also determines a preferred story style and story template $\mathcal{K}_{\texttt{tem}}$ based on the user profile $\mathcal{K}_{\texttt{usr}}$. In the NG stage, the framework first selects N key narrative elements, $\overline{G}_1, \overline{G}_2, \dots, \overline{G}_N$, based on the user’s query Q together with the $\mathcal{K}_{\texttt{dom}}$, $\mathcal{K}_{\texttt{tem}}$ and $\mathcal{K}_{\texttt{usr}}$. Each of the narrative element is a graph that consists of key actors, actions, and relations. Finally, a corresponding narrative, $S_{\texttt{nar}}$ and/or $V_{\texttt{nar}}$, is generated.

3.2 Processing Module

3.2.1 *Query Encoder.*

Queries are important inputs to the framework reflecting the users’ intention on which part of information is of interest and the form the narrative is expected to take. They are innately unstructured and multimodal, which calls for effective encoding and fusion methods. For textual queries, vectorised representation at word [55] or sentence level [45] have been proposed. Upon these representations, sequential models like hidden Markov model [12], recurrent networks [33, 73, 75] and Transformers [17, 81, 100] are used to model cross-word/sentence context. For speech inputs, audio features like frequency-based coefficients [68, 85] and spectrograms [40, 83] can be used as input to sequential models as mentioned above for context modelling. Video encoders like [7, 19, 79] focused on capturing local visual changes and aggregation of those information for global context modelling.

3.2.2 *Video Graph Generator.*

Video — unstructured temporal visual data — inherently contains a vast amount of semantic information, such as scenes [71], objects [53, 82, 89], fine-grained attributes [10], motion dynamic [96], relationship [47, 51], events [20], etc. The advances in network architecture [18, 31, 70] has achieved robust vision-based real-world sensing. Inspired by knowledge graphs, visual data can be represented as a graph [38, 42], in which each subject, attribute, and relation are denoted as node o and one (or two) edge is added to form a compositional concept (or relation triplets) such as $\langle\text{wet}, \text{dog}\rangle$ (or $\langle\text{man}, \text{carry}, \text{box}\rangle$). By parsing the raw video $\mathcal{V}$ into a video graph $\mathcal{G}$, the subsequent processing is agnostic to the input modality and allows the framework to freely integrate with different knowledge sources, *e.g.*, $\mathcal{K}_{\texttt{dom}}$ and $\mathcal{K}_{\texttt{tem}}$.

3.2.3 *Narrative Element Retrieval and Structure Modeling.*

The encoded query and generated video graph $\mathcal{G}$ are used as inputs to this module, where a retrieval process guided by external knowledge pinpoints the relevant elements in the video graph and story templates. Here, the narrative elements are the major events that compose a narrative. This is similar to cross-modal retrieval [105], activity retrieval [8], and video moment retrieval [76, 104, 106], but with external knowledge constraints. Specifically, the user query will determine what is the goal of the generated narrative, whereas other knowledge, such as user profile and domain knowledge, will assist the retrieval to identify more relevant narrative elements and allow the framework to handle unfamiliar topics. Once the narrative elements are retrieved, the system needs to analyse the causal relations, hierarchical relations, importance scores and the internal structures of these elements. The retrieved narrative elements serve as the factual material and as the input to the narrative generator.

3.2.4 *Narrative Generator.*

With all the information prepared by previous modules, this module finally assembles narrative elements, story templates, and domain knowledge into a narrative text $S_{\texttt{nar}}$ and an optional narrative video $V_{\texttt{nar}}$. As discussed in Section 2, the output narrative is expected to be: (a) correct as measured by natural language grammar and domain knowledge, and (b) interesting in terms of human aesthetics and user profile. While (a) is relatively easier as shown in existing storytelling literature, (b) requires a

Table 1: Comparison of narrative generation with related tasks. I: Image; V: Video; T: Text.

Task	Input Modality	Output Modality	Multi-hop Reasoning	Long Sentences	Generation	User Query	Storytelling	Causality
Visual Question Answer	I, V, T	T				✓		
Visual Dialogue	I, T	T	✓		✓	✓		
Image/Video Captioning	I, V	T	✓		✓			
Video Summary	V	V, T			✓			
Visual Storytelling	I, V, T	T		✓			✓	
Narrative Generation	V, T	V, T	✓	✓	✓	✓	✓	✓

more sophisticated integration of all the above-mentioned modules. Furthermore, there is no single ground truth narrative as there may exist multiple variants of narration that are both coherent and succinct. Hence, there is a need for research in computational narrative analysis [13] and narrative factual consistency [26, 43, 78].

3.3 External Knowledge

3.3.1 User Profile. The users' demographic information and interests are important sources from which the users' preference on the output narrative can be inferred. In the context of this work, the profile affects the user's expectation on the narrative: the factual contents it delivers, the arguments it proposes, the literature style it uses, etc. In recommendation systems literature [1, 32, 54, 108], it can be represented with respect to existing concepts, other users by tags [97], concept weights [14], or associative rules[21].

3.3.2 Story Templates. Story templates are schemes that control how narrative elements are arranged in the generated narrative, which could be explicitly defined as a cloze game, into which narrative elements are filled. It could also be defined as a set of heuristic rules, with which a planning algorithm, *e.g.*, IPOCL [63], discovers an arrangement of narrative elements. Interaction with humans [69] also helps the formation of a story. Recent story templates learning approach including learning the role transition of words from a text corpus to help generate a fluent story consisting of multiple sentences [60]. The storyline can also be distributed in the parameters of recurrent models [9, 93]. We aim at learning the story templates but from a broader perspective. The story templates are expected to not only guarantee logical correctness, but also ensure a aesthetically enjoyable narrative based on user profile. This requires learning different literature styles from a large text corpus, as well as take into account user query and other external knowledge.

3.3.3 Domain knowledge. A key factor that differentiates narrative generation from retrieval task is that it requires the model to go beyond similarity-based entity matching and achieve human-like knowledge-based reasoning. Domain knowledge is therefore an indispensable part of a successful NG framework. Existing knowledge databases provide domain knowledge using highly structured graphs or logic constraints. For example, Wikidata [77] provides common sense knowledge, while ConceptNet [74] provides taxonomy information between words. ATOMIC [37, 65] goes one step further by providing causal relationships between events, which serves as a domain knowledge for machine reasoning. All these knowledges form the basis of a variety of works on common sense-aware visual recognition [52, 88], image captioning [87, 99], VQA [58, 84], and neurosymbolic reasoning [24, 90].

3.4 Evaluation Metrics

3.4.1 Textual Factual Consistency Metric. The quality of narrative can be objectively validated via textual factual consistency, *i.e.*, RUBER [78] and FactCC [43]. The generated narrative is compared against the human narrated version, and a score between 0 to 1 indicates if the generated narrative has a better factual consistency. This ensures that the evaluation focus is on the narrative elements and not on the quality of the sentences. For the output video summary, one metric is to evaluate whether the output video covers the semantic content that is relevant to the text narrative.

3.4.2 Human Evaluation. Another aspect of the evaluation is the quality of style, fluency, and informativeness of the generated narrative (for both textual and visual output). Human evaluation is recommended as these factors are largely subjective, and the computational evaluation of story quality remains an open problem.

4 COMPARISON WITH RELATED TASKS

In this section, we review several closely related tasks and discuss the key differentiating factors to the proposed computational narrative generation task. The key differentiating factors are summarized in Table 1. Note that we neglect the perceptual computer vision task, such as semantic segmentation [30], detection and tracking [53, 98], object classification [18], person identification [16], relationship detection [47, 51, 91, 92], and so on, as those are the fundamental modules that serve as a building blocks for narrative generation.

First of all, the conventional multi-modal Question Answering (QA) task (*e.g.*, visual QA [2, 3, 41] and video QA [46, 109]) is formalized by a classification objective, with the aim of disentangling the capability of visual reasoning and text generation. As such, these tasks are limited by the inferior generalization in real open-ended scenarios. To address this, external knowledge is adapted to handle the open-domain scenario [64]. Another challenge for QA is the ambiguity in user's queries as well as the complexity of the multimedia data. To address this, visual dialogue [15, 61, 80] eases this pain via answering questions within multi-rounds. Nevertheless, the outputs are often expressed by a short sentence.

One may question the similarity between narrative generation and video captioning [50, 101, 107] or video summary [67, 102]. We argue that the proposed narrative generation is distinguished from these two tasks by four merits: 1) narrative generation is driven by an explicit purpose, which is goal-oriented and benefits from user queries; 2) it focuses mainly on long videos, as compared to short videos — often is a narrative by itself — in video captioning or video summary; 3) the causality is intrinsically ensured by the proposed narrative generation; and 4) narrative has four basic components, namely plot, character, conflict, and theme, and is required to be presented in a chronological order.

The task of storytelling has gained increasing attention [27, 35, 36, 48]. This include producing a story with a stack of images [35] or videos [48], and story ending generation with multi-sentence story plot [27, 36]. The targeted scope of these research sheds lights on the storytelling task and serve as a good foundation for the proposed narrative generation task. We argue that computational narrative generation has more inherent challenges, and therefore encourage the research community to approach the narrative generation task from multiple perspectives (*cf.* Section 3 and Section 5).

In a nutshell, narrative generation cannot be simply achieved by a simple composition of existing tools together due to the presence of these challenging inherent attributes – multi-hop reasoning, long sentences and long videos, user query, and causality.

5 RESEARCH PROBLEMS

This section discusses the unique challenges of narrative generation from three aspects, namely dataset, narrative construction, and narrative evaluation. These challenges are highly interdependent and could be considered either singly or jointly in the future work.

5.1 Dataset

5.1.1 Dataset Criterion. To facilitate the narrative generation studies, new benchmark datasets of long videos are needed. Existing datasets that satisfy this requirement include VideoSet [103] and Video Story [48]. Together, they consist of four TV episodes, 7 egocentric videos, and 105 videos from four types of common and complex events (*i.e.*, birthday, camping, Christmas and wedding). In future work, the key criteria to create a new dataset is to collect **long videos that has rich semantic events**. A common practice is to adopt narrated videos (*e.g.*, movies or TV series) or social events from video streaming platforms. However, we note that the narrated videos is by nature a narrative, and the selection of any such video should contain sufficiently complex story plots.

5.1.2 Dataset Annotation. There are three types of annotation that is required for the proposed task. (1) **Video Semantics**: A video graph $\mathcal{G} = (V, E)$ where V_i is a finite set of vertices and $E_{i,j}$ is an edge that connects V_i and V_j. The node including person, object, attribute, action, relationship, etc. (2) **Narrative Element and Structure**: While the common approach is to provide the textual summary of the given video [48, 66, 94, 103], we argue that a perfect narrative for each given query is hard to define. Hence, we suggest that for each query Q, a set of N key narrative elements $\overline{G}_1, \overline{G}_2, \ldots, \overline{G}_N$, *i.e.*, subgraphs in $\mathcal{G}$, that form a compelling story is annotated. In addition, structured information, such as the causal relationship, hierarchical relation among elements, importance score, plot, and so on, are also required. This will provide ground truth for factual consistency. (3) **Annotation Confidence**: As the annotators are likely not the domain experts, it would be important to allow them to indicate the confidence score for each annotated attributes. The confidence score should be released as the uncertainty of the annotation (*e.g.*, visual ambiguity) is an important feature that enables training of better models [49]. Another reason to indicate the confidence score is that the relations between elements in videos are implicit, but those in narratives are explicit (*cf.* Section 5.3.1).

5.2 Narrative Construction

5.2.1 Bridging Story Template, Contents, and Domain Knowledge. The narrative system we discussed above is a mixture of continuous (*e.g.*, encoded user query) and discrete information (*e.g.*, domain knowledge represented as knowledge graph). This has posed challenges about the integration of all the information into one narrative generation process. Intuitively, we would like the system to be end-to-end trainable to incorporate the massive multimedia data with little human intervention. This requires a flexible usage of both supervised and unsupervised learning approaches. On the other hand, embedding discrete knowledge into a continuous space requires neurosymbolic approaches as a bridge, which is itself an emerging field with unsolved issues.

5.2.2 Narrative Structure Modeling. A narrative is defined as a semiotic representation of a series of events connected in a temporal and causal way. Once the narrative elements $\overline{G}_1, \overline{G}_2, \ldots, \overline{G}_N$ are retrieved from the raw data or video graph, the system is required to infer the underlying order and manner (*i.e.*, the structural framework) that forms a narrative. The types of narrative structure include linear structure, non-linear structure, parallel structure and circular plot structure. The characters or events can also be connected via causal relation, conflict link [86], and hierarchical relation. In addition, the narrative elements also have different degrees of importance. The challenge in this research problem is to infer (1) how the narrative elements and nodes are interlinked with each other and (2) the corresponding attributes based on the selected narrative structure.

5.2.3 Personalized Narrative Generation. Although the factual narrative is based on the actual event with underlying temporal and causal relation, a good narrative is not bounded by a single objective story and enjoys various forms depending on the user's preference. Therefore, it is necessary to consider the user profile and preference, circumstances, and target audience during the narrative generation process. On one hand, such personalization can be based on user's preference on genre, style, or aesthetics. On the other hand, user's privacy concern would determine what elements can be included. For example, if the narrative is generated for family members, more private information can be included as opposed to news article. In addition, narrative persuasion would evoke various type of emotions. These factors have to be jointly considered during the narrative structure modeling and narrative generation.

5.3 Narrative Evaluation

Given the distinctive nature of narratives in our work, we face several new challenges in evaluation as discussed below.

5.3.1 Factual consistency. First of all, the evaluation of narrative completeness will be different from traditional video caption evaluations, as i) we focus on long (*e.g.*, spanning several days) videos and ii) our narrative is goal driven. Therefore, our definition of completeness will be more flexible based on various goals. In particular, our evaluation of completeness will need to consider all the following three aspects: i) whether all goal-related narrative elements are presented; ii) whether the temporal, causal and spatial relations of the narrative elements are fully described; iii) whether the story sufficiently serve the specific goals.

Correspondingly, we need to re-define narrative accuracy in the long video and goal-oriented scenario. Different from traditional metrics, our accuracy will be multidimensional. First, it measures the content's consistency between narratives and videos. Second, it evaluates the correlation among narrative descriptions and various goals (*e.g.*, if the mentioned stories serve each goal). Finally, it should be able to assess the fidelity of the narrative elements structure. For example, the causal relations in videos are often implicit and subjective to human perceptions. Some causal relations are obvious and definite, but others may be subtle and ambiguous. While transferring such relations to explicit narratives, accuracy measures how faithfully such relations are described: "Event A causes event B to happen" represents a causal relation of high confidence, whereas "Event B happens after event A" indicates an obscure cause-effect.

5.3.2 *High-level Narrative Attribute Assessment.*

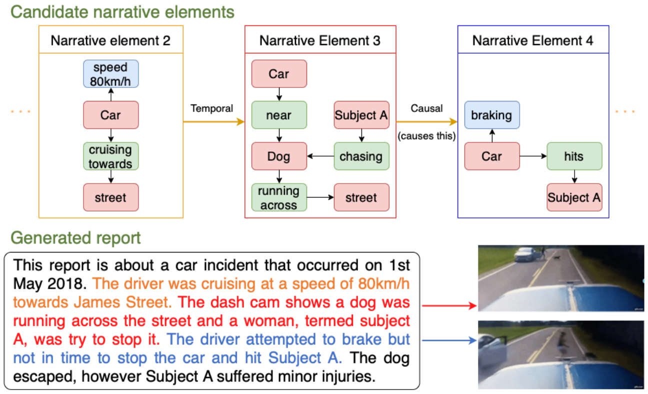
Different from traditional video captioning, our narrative is goal driven. This means that it should be able to meet specific requirements from the users, and fulfill its task of narrative persuasion.

Therefore, besides completeness and accuracy, a good narrative in our study should have certain high-level properties, such as being coherent, fluent, informative yet compact, and amiable. In some applications like smart stores (*e.g.*, Amazon Go), privacy-preserving may be a key requirement for the narratives. The above attributes are often implicit and subject, yet they are vital contributing factors to narrative persuasion. Notably, the attributes may be inter-correlated and even competing. For example, sometimes we need to consider the trade-offs between informative and privacy-preserving. All the above create new and tough challenges for the high-level assessment of narratives. One possible way to tackle this is to design a subjective evaluation standard for each attribute, similar to the mean opinion score (MOS) evaluation for audio [59].

6 CONCEPTUAL APPLICATION

Motivated by the real-world impact of narrative generation (NG), we illustrate how NG can be applied in Video Logging as a Service (VLaaS). Existing commercial VLaaS applications, such as road condition monitoring[1] and geographic information systems[2], provide the data capture and storage solution, advanced analytic, and a customized interface that allow its clients to access the data. Another akin service is Google Photos[3] where utilities like photo search, person album management, and memories generation are provided.

In contrast, NG aims to generate a text narrative with supporting video evidences. The user can provide verbal command with a preferred style and target audience, and the VLaaS platform will handle it automatically. In the following, we provides three NG use cases. The first use case is Productivity Monitoring, such as in smart manufacturing or smart retail. This is aligned with Industry 4.0 where big data are combined with advanced AI models to generate informative reports. For example, the factory can use the long term causal inference to understand why the production quality has dropped. On the other hand, footage from the retail environment can be used to understand how the marketing of new product affects the customer engagement and sales. The second use case is Legal Report Generation, where video footage from car's dashcam, smart door bell or home monitoring camera can be used to generate a insurance claims report that highlight the details of an incident. An example of car accident report is shown in Fig. 3. The third use case is Unmanned Region Monitoring, where smart cameras or drones can be deployed over unmanned and remote areas (*e.g.*, forests, oceans, volcanic regions, etc.) The collected data can be used to investigate how a forest fire has spread, or to understand the animal behaviour so the scientist can better preserve the endangered population. Ultimately, the holy grail of narrative generation in this use case is to generate a National Geographic style article without human editing.



This report is about a car incident that occurred on 1st May 2018. The driver was cruising at a speed of 80km/h towards James Street. The dash cam shows a dog was running across the street and a woman, termed subject A, was try to stop it. The driver attempted to brake but not in time to stop the car and hit Subject A. The dog escaped, however Subject A suffered minor injuries.

Figure 3: An example of incident report generation for insurance claim. The goal here is to generate an incident report of an accident, and the input is the dashcam recording from the policy holder's vehicle.

7 CONCLUSION

Computational goal-driven narrative generation with long videos is a challenging and ambitious problem in multimedia research. We present a particular use case of factual narrative, where events captured in multiple long videos are analyzed and connected in a temporal and causal way. The paper introduces a general framework as the basis to discuss the research problems and challenges, as well as compare it with related tasks in the multimedia literature.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

[1] http://www.dcl.co.nz/services/video-logging
[2] https://support.esri.com/en/technical-article/000024386
[3] https://photos.google.com

REFERENCES

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2011. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *UMAP (Lecture Notes in Computer Science, Vol. 6787)*. Springer, 1–12.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*. 6077–6086.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*. 2425–2433.

[4] Markus Appel and Tobias Richter. 2007. Persuasive effects of fictional narratives increase over time. *Media Psychology* 10, 1 (2007), 113–134.

[5] Skylar Bayer and Annaliese Hettinger. 2019. Storytelling: A Natural Tool to Weave the Threads of Science and Community Together. *The Bulletin of the Ecological Society of America* 100, 2 (2019), e01542.

[6] Sébastien Caquard and Daniel Naud. 2014. A spatial typology of cinematographic narratives. *Modern Cartography Series* 5 (2014), 161–174.

[7] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*. 4724–4733.

[8] Gregory D. Castañón, Yuting Chen, Ziming Zhang, and Venkatesh Saligrama. 2015. Efficient Activity Retrieval through Semantic Graph Queries. In *ACM Multimedia*. 391–400.

[9] Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. Commonsense Knowledge Aware Concept Selection For Diverse and Informative Visual Storytelling. In *AAAI*. 999–1008.

[10] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-Grained Video-Text Retrieval With Hierarchical Graph Reasoning. In *CVPR*. 10635–10644.

[11] Dorrit Cohn. 2000. *The distinction of fiction*. JHU Press.

[12] John M. Conroy and Dianne P. O'Leary. 2001. Text Summarization via Hidden Markov Models. In *SIGIR*. 406–407.

[13] Martin Cortazzi. 1994. Narrative analysis. *Language Teaching* 27, 3 (1994), 157–170.

[14] Mariam Daoud, Lynda Tamine-Lechani, Mohand Boughanem, and Bilal Chebaro. 2009. A session based personalized search using an ontological user profile. In *SAC*. 1732–1736.

[15] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *CVPR*. 1080–1089.

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*. 4690–4699.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *ICCV*. 6201–6210.

[20] Tian Gan, Yongkang Wong, Daqing Zhang, and Mohan S. Kankanhalli. 2013. Temporal encoded F-formation system for social interaction detection. In *ACM Multimedia*. 937–946.

[21] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. 2007. User Profiles for Personalized Information Access. In *The Adaptive Web (Lecture Notes in Computer Science, Vol. 4321)*. Springer, 54–89.

[22] James Paul Gee and Francois Grosjean. 1984. Empirical evidence for narrative structure. *Cognitive Science* 8, 1 (1984), 59–85.

[23] John C Georgesen and Cecilia H Solano. 1999. The effects of motivation on narrative content and structure. *Journal of Language and Social Psychology* 18, 2 (1999), 175–194.

[24] Eleonora Giunchiglia and Thomas Lukasiewicz. 2021. Multi-Label Classification Neural Networks with Hard Logical Constraints. *Journal of Artificial Intelligence Research* 72 (2021), 759–818.

[25] Melanie C Green. 2006. Narratives and cancer communication. *Journal of Communication* 56 (2006), S163–S183.

[26] Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In *EMNLP*. 9157–9166.

[27] Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story Ending Generation with Incremental Encoding and Commonsense Knowledge. In *AAAI*. 6473–6480.

[28] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan S. Kankanhalli. 2018. Multi-modal Preference Modeling for Product Search. In *ACM Multimedia*. 1865–1873.

[29] Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan Kankanhalli. 2022. A Unified End-to-End Retriever-Reader Framework for Knowledge-based VQA. In *ACM Multimedia*.

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2020. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2 (2020), 386–397.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[32] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. 173–182.

[33] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[34] Hans Hoeken and Jop Sinkeldam. 2014. The role of identification and perception of just outcome in evoking emotions in narrative persuasion. *Journal of Communication* 64, 5 (2014), 935–955.

[35] Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Oliver Wu, Jianfeng Wang, and Xiaodong He. 2019. Hierarchically Structured Reinforcement Learning for Topically Coherent Visual Story Generation. In *AAAI*. 8465–8472.

[36] Qingbao Huang, Chuan Huang, Linzhang Mo, Jielong Wei, Yi Cai, Ho-fung Leung, and Qing Li. 2021. IgSEG: Image-guided Story Ending Generation. In *ACL/IJCNLP (Findings)*. 3114–3123.

[37] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *AAAI*. 6384–6392.

[38] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs. In *CVPR*. 10233–10244.

[39] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*. 3128–3137.

[40] Gibak Kim, Yang Lu, Y. Hu, and Philipos C. Loizou. 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America* 126, 3 (2009), 1486–1494.

[41] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 5583–5594.

[42] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[43] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *EMNLP*. 9332–9346.

[44] Jose Angel Garcia Landa. 2005. Narrative theory. *University of Zaragoza. On Line Edition* (2005).

[45] Quoc V. Le and Tomás Mikolov. 2014. Distributed Representations of Sentences and Documents. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 32)*. JMLR.org, 1188–1196.

[46] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Action-Centric Relation Transformer Network for Video Question Answering. In *CVPR*. 9972–9981.

[47] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2017. Dual-Glance Model for Deciphering Social Relationships. In *ICCV*. 2669–2678.

[48] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2020. Video Storytelling: Textual Summaries for Events. *IEEE Transactions on Multimedia* 22, 2 (2020), 554–565.

[49] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2020. Visual Social Relationship Recognition. *International Journal of Computer Vision* 128, 6 (2020), 1750–1764.

[50] Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. 2022. Long Short-Term Relation Transformer With Global Gating for Video Captioning. In *IEEE Transactions on Image Processing*, Vol. 31. 2726–2738.

[51] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual Relationship Detection with Language Priors. In *ECCV (Lecture Notes in Computer Science, Vol. 9905)*. Springer, 852–869.

[52] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The More You Know: Using Knowledge Graphs for Image Classification. In *CVPR*. 20–28.

[53] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. 2022. TrackFormer: Multi-Object Tracking with Transformers. In *CVPR*. 8844–8854.

[54] Manel Mezghani, Corinne Amel Zayani, Ikram Amous, and Faïez Gargouri. 2012. A user profile modelling using social annotations: a survey. In *WWW (Companion Volume)*. 969–976.

[55] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 3111–3119.

[56] Emily Moyer-Gusé and Robin L Nabi. 2010. Explaining the effects of narrative in an entertainment television program: Overcoming resistance to persuasion. *Human communication research* 36, 1 (2010), 26–52.

[57] J Murphy, S McDonough, R van Haren, B Triglone, and J Salinas. 2001. Practical strategies: STELLA narratives. *Literacy Learning: the Middle Years* 9, 2 (2001).

[58] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 2018. Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question

Answering. In *NeurIPS*. 2659–2670.

[59] Ira L Panzer, Alan D Sharpley, and William D Voiers. 1993. A comparison of subjective methods for evaluating speech quality. In *Speech and audio coding for wireless and network applications*. Springer, 59–65.

[60] Cesc C. Park and Gunhee Kim. 2015. Expressing an Image Stream with a Sequence of Natural Sentences. In *NIPS*. 73–81.

[61] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two Causal Principles for Improving Visual Dialog. In *CVPR*. 10857–10866.

[62] Brian Richardson. 2000. Recent concepts of narrative and the narratives of narrative theory. *Style* 34, 2 (2000), 168–175.

[63] Mark O. Riedl and Robert Michael Young. 2010. Narrative Planning: Balancing Plot and Character. *Journal of Artificial Intelligence Research* 39 (2010), 217–268.

[64] Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering. In *NeurIPS*.

[65] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI*. 3027–3035.

[66] Anna Senina, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. 2014. Coherent Multi-Sentence Video Description with Variable Level of Detail. In *German Conference on Pattern Recognition*. 184–195.

[67] Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal Video Summarization via Time-Aware Transformers. In *ACM Multimedia*. 1756–1765.

[68] Yang Shao and DeLiang Wang. 2008. Robust speaker identification using auditory features and computational auditory scene analysis. In *ICASSP*. 1589–1592.

[69] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. 2021. Calliope: Automatic Visual Data Story Generation from a Spreadsheet. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 453–463.

[70] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[71] Josef Sivic and Andrew Zisserman. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*. 1470–1477.

[72] Jeffrey R. Smitten and Ann Daghistany. 1981. *Spatial Form in Narrative*. Cornell Univ Press.

[73] Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. 2011. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In *ICML*. 129–136.

[74] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*. 4444–4451.

[75] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*. 3104–3112.

[76] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A. Plummer. 2021. LoGAN: Latent Graph Co-Attention Network for Weakly-Supervised Video Moment Retrieval. In *WACV*. 2082–2091.

[77] Thomas Pellissier Tanon, Denny Vrandecic, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From Freebase to Wikidata: The Great Migration. In *WWW*. 1419–1428.

[78] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *AAAI*. 722–729.

[79] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *CVPR*. 6450–6459.

[80] Tao Tu, Qing Ping, Govindarajan Thattai, Gökhan Tür, and Prem Natarajan. 2021. Learning Better Visual Dialog Agents With Pretrained Visual-Linguistic Representation. In *CVPR*. 5622–5631.

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[82] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. 2019. MOTS: Multi-Object Tracking and Segmentation. In *CVPR*. 7942–7951.

[83] DeLiang Wang and Jitong Chen. 2018. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26, 10 (2018), 1702–1726.

[84] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2018. FVQA: Fact-Based Visual Question Answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 10 (2018), 2413–2427.

[85] Yuxuan Wang, Kun Han, and DeLiang Wang. 2013. Exploring Monaural Features for Classification-Based Speech Segregation. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 21, 2 (2013), 270–279.

[86] Stephen G. Ware, R. Michael Young, Brent Harrison, and David L. Roberts. 2014. A Computational Model of Plan-Based Narrative Conflict at the Fabula Level. *IEEE Transactions on Computational Intelligence and AI in Games* 6, 3 (2014), 271–288.

[87] Shuang Wu, Shaojing Fan, Zhiqi Shen, Mohan S. Kankanhalli, and Anthony K. H. Tung. 2020. Who You Are Decides How You Tell. In *ACM Multimedia*. 4013–4022.

[88] Shuang Wu, Mohan S. Kankanhalli, and Anthony K. H. Tung. 2022. Superclass-aware network for few-shot learning. *Computer Vision and Image Understanding* 216 (2022), 103349.

[89] Yu Xiang, Alexandre Alahi, and Silvio Savarese. 2015. Learning to Track: Online Multi-object Tracking by Decision Making. In *ICCV*. 4705–4713.

[90] Yaqi Xie, Ziwei Xu, Kuldeep S. Meel, Mohan S. Kankanhalli, and Harold Soh. 2019. Embedding Symbolic Knowledge into Deep Networks. In *NeurIPS*. 4235–4245.

[91] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2020. Interact as You Intend: Intention-Driven Human-Object Interaction Detection. *IEEE Transactions on Multimedia* 22, 6 (2020), 1423–1432.

[92] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. 2019. Learning to Detect Human-Object Interactions With Knowledge. In *CVPR*. 2019–2028.

[93] Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu. 2021. Imagine, Reason and Write: Visual Storytelling with Graph Knowledge and Relational Reasoning. In *AAAI*. 3022–3029.

[94] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*. 5288–5296.

[95] Ning Xu, An-An Liu, Yongkang Wong, Yongdong Zhang, Weizhi Nie, Yuting Su, and Mohan S. Kankanhalli. 2019. Dual-Stream Recurrent Neural Network for Video Captioning. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 8 (2019), 2482–2493.

[96] Ziwei Xu, Xudong Shen, Yongkang Wong, and Mohan S. Kankanhalli. 2021. Unsupervised Motion Representation Learning with Capsule Autoencoders. In *NeurIPS*. 3205–3217.

[97] Su Yan, Xin Chen, Ran Huo, Xu Zhang, and Leyu Lin. 2020. Learning to Build User-tag Profile in Recommendation System. In *CIKM*. 2877–2884.

[98] Fan Yang, Xin Chang, Sakriani Sakti, Yang Wu, and Satoshi Nakamura. 2021. ReMOT: A model-agnostic refinement for multiple object tracking. *Image and Vision Computing* 106 (2021), 104091.

[99] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. 2020. Hierarchical Scene Graph Encoder-Decoder for Image Paragraph Captioning. In *ACM Multimedia*. 4181–4189.

[100] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*. 5754–5764.

[101] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing Videos by Exploiting Temporal Structure. In *ICCV*. 4507–4515.

[102] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. In *CVPR*. 982–990.

[103] Serena Yeung, Alireza Fathi, and Li Fei-Fei. 2014. In VideoSET: Video Summary Evaluation through Text. *CVPR Workshop*.

[104] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval. In *CVPR*. 2215–2224.

[105] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2013. Heterogeneous Metric Learning with Joint Graph Regularization for Cross-Media Retrieval. In *AAAI*. 1198–1204.

[106] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. 2019. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. In *CVPR*. 1247–1257.

[107] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020. Object Relational Graph With Teacher-Recommended Learning for Video Captioning. In *CVPR*. 13278–13288.

[108] Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. 2011. Functional matrix factorizations for cold-start recommendation. In *SIGIR*. 315–324.

[109] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. 2017. Uncovering Temporal Context for Video Question and Answering. In *International Journal of Computer Vision*, Vol. 124. Springer, 409–421.