Computational Visual Media
https://doi.org/10.1007/s41095-0xx-xxxx-x

Research/Review Article

Script-to-Storyboard: A New Contextual Retrieval Dataset and Benchmark

Xi Tian[1] (✉), **Yong-Liang Yang**[1], **and Qi Wu**[2]

© The Author(s)

Abstract Storyboards comprising key illustrations and images help filmmakers to outline ideas, key moments, and story events when filming movies. Inspired by this, we introduce the first contextual benchmark dataset Script-to-Storyboard (Sc2St) composed of storyboards to explicitly express story structures in the movie domain, and propose the contextual retrieval task to facilitate movie story understanding. The Sc2St dataset contains fine-grained and diverse texts, annotated semantic keyframes, and coherent storylines in storyboards, unlike existing movie datasets. The contextual retrieval task takes as input a multi-sentence movie script summary with keyframe history and aims to retrieve a future keyframe described by a corresponding sentence to form the storyboard. Compared to classic text-based visual retrieval tasks, this requires capturing the context from the description (script) and keyframe history. We benchmark existing text-based visual retrieval methods on the new dataset and propose a recurrent-based framework with three variants for effective context encoding. Comprehensive experiments demonstrate that our methods compare favourably to existing methods; ablation studies validate the effectiveness of the proposed context encoding approaches.

Keywords Dataset, benchmark, text-based-image retrieval, movie

1 Introduction

Movies, one of the most complex visual art forms, simulate experiences that communicate ideas, stories, perceptions, feelings, beauty, or atmosphere through the use of moving images. In recent years, there has been increasing computing research focusing on various aspects of movies, such as movie shot retrieval [1], action recognition [2], and question answering [3]. Also, many movie domain datasets [2–8] have been proposed to facilitate movie understanding. These datasets are getting larger, and often include various annotations including subtitles, plots, descriptions, and so on.

Despite the proliferation of movie-related datasets, far too little attention has been paid to the story structure in movies: the way they express the stories.

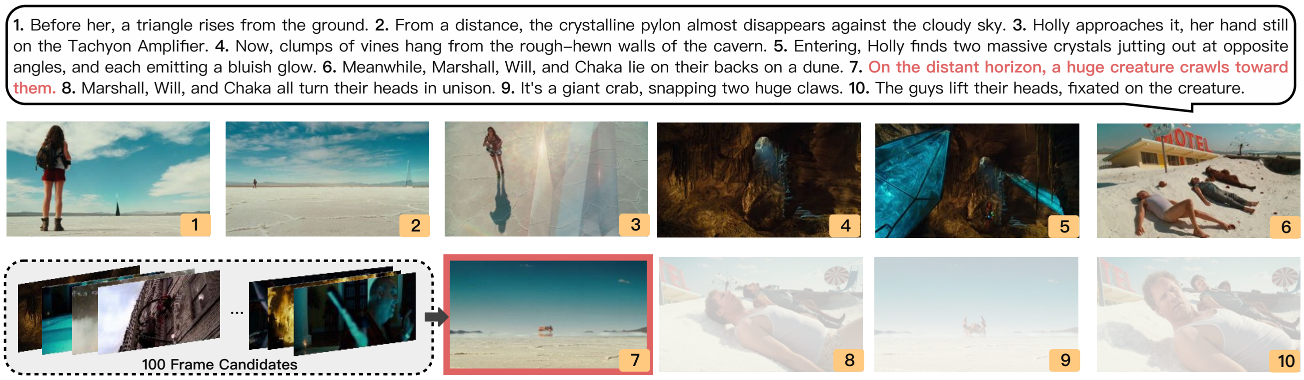
Naturally, videos are the carrier of movie stories. Stories are, however, implicitly contained in the video consisting of shots, visual elements, sounds and dialogues, *etc*. We argue that video clips, as the final format of film making, usually mix multiple storylines, making the story structure complex and unclear. For example, the minimum video clip of many existing movie datasets [3, 7] is at the minute level. Likewise, the corresponding text descriptions are highly generalized (like plots or synopses). Long videos and high-level text make optimization difficult in certain tasks. Some datasets tend to use cartoons to highlight the story structure [9, 10], which favours simple and clear story content, but in this way, they ignore the real-world visual elements. We agree with the benefits of a clear story structure, and in this way, it would be more suitable as a testbed for movie story understanding.

Actually, stories in movies can be explicitly expressed using the *storyboard*, a graphic layout of sequential illustrations and images to visually tell a story. Storyboards are useful in filming movies to express the key moments and outline the events. The usage of storyboards shows that movie clip videos can be condensed to *keyframes*. Inspired by this, we aim to construct a storyboard-based dataset to clearly outline story structure to facilitate story understanding. We name the dataset Script-to-Storyboard (Sc2St), along with which we introduce a new processing task in the movie domain: contextual text-based-image retrieval. Fig. 1 shows a storyboard sample and the retrieval task. A storyboard in our work is composed of a series of still video frames, with drawings/pictures of sequential key events and shots in a film.

1 Department of Computer Science, University of Bath, Bath BA2 7AY, United Kingdom. E-mail: X. Tian (✉), xt275@bath.ac.uk; Y. Yang, y.yang@cs.bath.ac.uk.

2 Australian Institute for Machine Learning, School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia. E-mail: Q. Wu, qi.wu01@adelaide.edu.au.

Manuscript received: 2022-01-01; accepted: 2022-01-01



1. Before her, a triangle rises from the ground. **2.** From a distance, the crystalline pylon almost disappears against the cloudy sky. **3.** Holly approaches it, her hand still on the Tachyon Amplifier. **4.** Now, clumps of vines hang from the rough–hewn walls of the cavern. **5.** Entering, Holly finds two massive crystals jutting out at opposite angles, and each emitting a bluish glow. **6.** Meanwhile, Marshall, Will, and Chaka lie on their backs on a dune. **7.** On the distant horizon, a huge creature crawls toward them. **8.** Marshall, Will, and Chaka all turn their heads in unison. **9.** It's a giant crab, snapping two huge claws. **10.** The guys lift their heads, fixated on the creature.

Fig. 1 Contextual retrieval task using the Script-to-Storyboard dataset. A storyboard's future keyframes are retrieved one by one, given a movie summary script comprising multiple sentences that depict the whole story, the keyframe history, and the sentences for future frames. Each round of retrieval selects from a list of frame candidates.

In the Sc2St task, we take a multi-sentence paragraph (or *script*, for short) as the story description. Then, for each movie frame, given its query text, the task asks a model to retrieve the best match from a prepared list of candidates for the current frame. These are carefully collected in-movie and cross-movie frames, serving as a benchmark for evaluation of the retrieval performance. Section 4.1 defines the task in detail, and the evaluation setting.

This task differs from conventional text-based retrieval, such as text-based-image and text-based-video retrieval. Text-based-image retrieval uses images as visual content, but it only considers independent text-based-image pairs without considering the additional context. Text-based-video retrieval uses video clips as visual content. Although videos contain frames used as the in-clip context, the dense neighbouring frames look about the same, resulting in redundant visual information and unclear contextual structure. Actually, the text-based-video retrieval task also ignores across-video context, making it similar to text-based-image retrieval except for the usage of clips rather than images. There are deeper challenges in the Sc2St task: in a storyboard, a future keyframe is not only related to its textual caption, but also the script telling the story; it should also be visually coherent with previous keyframes. Existing context-aware image retrieval methods focus on modelling the context only from the text [11, 12], *e.g.*, building the textual context from sentence level to paragraph level. In comparison, the Sc2St task requires a model to capture contextual information from both visual and textual streams.

In constructing the Sc2St dataset, we have taken into account that there are already sufficient datasets in the movie domain, such as [2–8]. To avoid repeated annotations, we have selected the LSMDC [6] movie collection as a basis and provide further fine-grained text and keyframe annotations to it. We processed all available 165 movies in three steps. Firstly, we produced fine-grained text by systematically parsing the original video-level text descriptions into sentences. Then, using rule-based methods to recognize 'splittable' sentences, we further split those into sub-sentences to generate finer and elementary text instances. Secondly, we identified keyframes by selecting the most representative images from the video clips for each text instance using similarity-based and manual screening, take into account both objective and subjective factors. Thirdly, we formed stories as storyboards containing a sequence of text-keyframe pairs using keyframes temporally close in the original, to ensure semantic storyline coherence. The storyboard length is by default set to ten items in our work but is flexible to length. We finally collated the Sc2St dataset consisting of ≈20.4k storyboards with ≈61.2k unique keyframes and ≈44k unique text instances (sentences or sub-sentences describing the keyframe). Statistical analysis reveals that the newly created Sc2St dataset has several advantages: (i) finer-grained text than existing movie datasets, (ii) highly-diverse textual descriptions, and (iii) better story coherency than in story understanding datasets. To tackle the contextual retrieval task based on the new dataset, we propose a recurrent model framework with three variants targeting dual-way context encoding. We have conducted extensive experimental comparisons to existing text-based-image and text-based-video retrieval approaches. We further evaluate human performance using an online testing system for comparison. We also provide detailed analyses and discussions to demonstrate the effectiveness of our model in capturing the contextual information in the new task.

Our main contributions are in summary:

(1) a new benchmark dataset Sc2St in the movie domain with novel story structures: it comprises storyboards with coherent story structures, fine-grained texts, and semantic keyframes,
(2) the contextual retrieval task (script-to-storyboard): given a script, it aims to retrieve future keyframes by respecting both the corresponding text and image history coherence,
(3) baseline methods for this task, with extensive experimental comparison to existing methods and human performance, and
(4) discussions of the effectiveness of our methods and of the potential of our dataset for image generation.

2 Related Work

2.1 Related Datasets

The proposed dataset has two features, where the contents are in the *movie* domain and the storyboard form relates to *story understanding*. We thus present the related datasets with analysis from the perspective of movie datasets and datasets involving sequential story-based tasks.

2.1.1 Movie Datasets

MovieQA [3] was collected from 408 movies and targets story and video understanding using question-answering. The data include video clips, plots, subtitles, scripts, *etc*. However, there are some concerns about using MovieQA for story understanding: video clips are not always provided, the text sources vary largely in detail and are not rich in content, the video clips are minutes long, lacking fine-grained timestamps, *etc*. Going beyond MovieQA, we provide a keyframe-based story dataset with fine-grained annotation. *MovieGraphs* [4] provides graph-based annotations of detailed social situations from 51 movies. The graph consists of various nodes capturing the characters' presence, their emotional and physical attributes, their relationships and interactions, *etc*. MovieGraph focuses mainly on using graphs for movie situation recognition, while our work targets contextual movie story understanding using textual and visual data. *AVA* [2] is an action recognition dataset sourced from 430 movies with annotations including 80 atomic visual actions in space and time on 15-minute video clips. Although the AVA dataset has densely labelled person-centric actions, the clips used are only part of the original movies and the action information is not rich enough for story understanding compared to use of textual descriptions. *MSA* [5] contains 327 movies for movie story understanding in matching between movie segments and synopsis paragraphs. It gives each movie a synopsis and provides annotated associations between corresponding synopsis paragraphs and movie segments. Although MSA also splits each movie segment into multiple shots (events), unlike our fine-grained text and keyframe pairs, they MAY match one paragraph of synopsis to many movie shots due to the high-level descriptive nature of the movie synopsis. *LSMDC* [6] consists of nearly 128k video clips annotated with detailed descriptive sentences from around 200 movies. The textual descriptions are collected from the transcribed audio description (AD), which gives a descriptive narration of important visual elements of movie clips for visually impaired people. The usage of AD ensures that the textual description captures the key story as well as the necessary details, unlike high-level synopsis/plots or redundant subtitles with less visual narrative. Our proposed dataset is based on LSMDC, while the differences are: (i) we use *semantic keyframes* rather than videos as visual forms, (ii) we further prepare the fine-grained text-keyframe pairs by aligning each sub-sentence with a key image, and (iii) we introduce additional character and meta-annotations to enrich the existing dataset. *Condensed Movies* [8] targets long-range understanding of the narrative structures of movies. It consists of around 36k movie key scenes with high-level descriptions and character face tracks. The collected movie segments are freely available from YouTube and the number of movies involved is much larger than for other movie datasets. However, as each movie includes roughly ten segments (the *key scenes*), the clip duration is long while the description is short. Learning the relationship between coarse text and complex video is difficult. The recent *MovieNet* [7] is a holistic dataset for movie understanding. It was sourced from 1.1k movies, comprising annotations of movie trailers, photos, plot descriptions, character information, scene boundaries, descriptions, *etc*. Despite the various, large-scale annotations, this dataset shares a problem with Condensed Movies: the aligned movie segments and descriptions are at a high level, *i.e*., the text source is a synopsis paragraph and the clips are up to a few minutes.

2.1.2 Story Datasets

PororoQA [9] focuses on video story question-answering on cartoon videos, with around 16k scene-dialogue pairs. The dialogues contain fine-grained sentences for scene descriptions, so the Pororo dataset contains not only rich descriptive details but also simple and coherent story structures. Related research also uses Pororo for story-based image generation [13]. However, the cartoon domain restricts the diversity of genres, scenes, and characters compared to the hundreds of movies in our dataset. *CoDraw* [10] is a collaborative image-drawing game that contains visual and movable clip art objects. The game has two players communicating together to construct a scene: a teller describes an abstract scene while the drawer reconstructs the scene or asks for details. The collected dataset contains ≈10K dialogues with corresponding scenes in multiple rounds. CoDraw is similar to our dataset since each round's scene image is like a movie keyframe and each dialogue tells a coherent story. However, unlike our movie-based dataset, CoDraw is based on cartoon art where the visual elements are simple and the involved actions are predefined. The *Visual Storytelling Dataset* [14] was proposed with the aim of generating image sequences from language. This dataset has a similar structure to ours: consecutive images, each provided with a corresponding description. However, the data are collected from the Internet which constrains the forming of image sequences, capable of being turned into a story: for example, some topics like 'birthday' or 'party' were manually pre-defined. Also, some images are missing due to deletion by posters.

2.2 Related Methods

Visual-language retrieval is the most closely related area, particularly text-based-image and text-based-video retrieval; we also investigate retrieval tasks involving sequential encoding. Text-to-image generation tasks are also discussed since our Sc2St dataset has the potential for sequential image generation.

2.2.1 Text-based Image Retrieval

Existing works focus on visual-semantic embedding for learning the similarities between the two modalities. Some methods have been proposed to improve the ranking losses used, such as exploiting hardest negative pairs [15] or instance losses [16] to improve the discriminative representation, and projection classification loss [17] to categorize one modality representation vector to another. Other methods further employ fine-grained image-sentence matching [18, 19], *e.g.*, matching words with image regions. Recently, the pre-training-based transformers [20] have become more popular, achieving significant performance gains in multiple tasks. Representative visual-language transformers include Uniter [21], and Oscar [22], which leverage word-region alignment to learn the image-text representations.

2.2.2 Text-based Video Retrieval

A widely used approach in video-language retrieval is to learn a joint embedding space from similar texts and videos [23]. Recent state-of-the-art methods [23, 24] follow the mixture-of-experts (MoE) paradigm [25] by combining several different embeddings from pre-trained models for video representation. Videos are composed of images or frames and can be taken as single-frame form using the image features in analysis [26], so the MoE framework can be adopted in our task.

2.2.3 Contextual Retrieval

In the retrieval tasks, contextual retrieval is a type that considers either the context in the structure of the text or the images/videos. The context usually exists in sequential data forms, *e.g.*, the sequence of sentences in the text or the continuous frames in the videos. [11] proposes a hierarchical (from sentence to story) text encoder to encode each sentence and retrieve the relevant images. [12] aims to create the storyboard by combining retrieval and style transfer. Its story-to-image retriever also uses a hierarchical text encoding method, *e.g.*, from word to sentence to story. However, these methods capture the context only from the text part. The most similar to ours is the Contextual Mixture of Embedding Experts model (CMoEE) [8] which adds context both from past and future movie clips to learn the text-based-video similarity. The used experts not only include the movie scene/objects representations but also the character embeddings. However, there is no textual context modelling or global story description constrained here.

2.2.4 Text to Visual Content Generation

Reed *et al.* [27] first proposed the text-to-image (T2I) model using conditional generative adversarial networks (GANs). Afterwards, other research made various improvements, e.g. to *image quality* by using coarse-to-fine structure [28, 29], *text-based-image consistency* based on an attention mechanism [30–32], *etc*. Recently, large-scale T2I models have brought remarkable advances in realistic image synthesis [33–35]. Instead of generating a single image, [13] can synthesize a series of cartoon images using a recurrent-based generative model. Similarly, [36] iteratively generates images from continual linguistic instructions at multiple steps.

Some applications allow taking intuitive user input, *e.g.*, paragraphs or multiple sentences, for content creation [37], such as text-guided storytelling [14] and video editing [38]. [14] focuses on sequential image retrieval, while [38] can create video montages made from retrieved video shots based on user-specified texts. These applications share a similar sequential retrieval form with Sc2St. However, the aims differ. The Sc2St dataset naturally contains movie story context and is built with the objective of contextual retrieval analysis. In contrast, [14, 38] use general topics (tour, party, or animals) to create the video context with the objective of content creation.

Table 1 Characteristics of various datasets.

Dataset	Domain	Textual Annotation	Aligned visual unit	Character	Coherency types
AVA	movie	actions	-	✓	-
MovieGraphs	movie	graphs, description	long clips	✓	graphs
MovieQA	movie	subtitles, plots, scripts	segments	✓	-
MSA	movie	plots	segments	✗	-
LSMDC	movie	AD, scripts	short clip	✗	-
Condensed Movies	movie	high-level description	segments	✓	-
MovieNet	movie	script, synopsis, subtitles, plots	segments	✓	-
PororoQA	cartoon	description	short clips	✓	video
CoDraw	cartoon	dialogues	keyframes	✓	sequential images
Visual Storytelling	open	description	keyframes	✗	sequential images
Sc2St	movie	(sub-)sentences, meta, actions	keyframes	✓	sequential images

Table 2 Statistical analysis of various datasets. N/A: information not available. *: data computed from derived video clips.

Dataset	#videos	#visual units	avg. unit dur.(s)	#sents / unit	#words / sent	dur.(s) / sent
AVA	430	-	-	-	-	-
MovieGraphs	51	7.6k	44.28	2.73	12.9	16.2
MovieQA	140	6.8k	202	N/A	N/A	N/A
MSA	327	4.5k	413.3	5.9	21.8	70.0
LSMDC	204	128k	4.1	1.0	9.0	4.1
Condensed Movies	3605	334k	134	-	-	-
MovieNet	1100	4.2k	428	5.9	-	72.5
PororoQA	171	16k	4.6	2.7	-	1.7
CoDraw	-	70k	-	1.97	16	-
Visual Storytelling	-	81.7k	-	1.07	11.4	-
Sc2St	165	61.2k	4.25*	1.08*	10.9	3.94*

3 The Sc2St Dataset

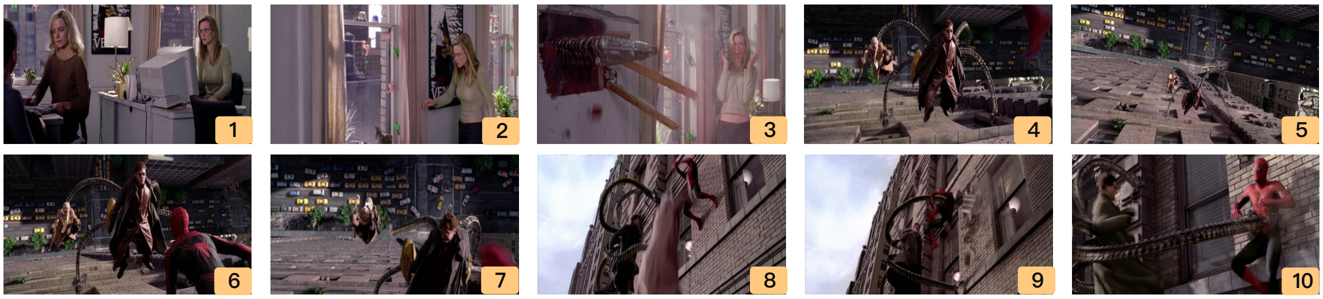
3.1 Dataset Construction

The successful usage of storyboards in film making reveals their power to illustrate the condensed story. The sequential structure explicitly delivers a more concise but coherent visual story structure, and we thus aim to use a storyboard as the data form. Similar image-based story datasets include CoDraw [10] and Visual Storytelling [14]. However, [10] uses synthetic cartoon data that is limited in scene and character variety and lacks the generality of the real world. [14] collects image-text pairs independently from the Internet and constructs stories manually; as a result, the story cohesion may not be strong and neighbouring image styles are usually different. Thus, we choose the movie domain because movies not only naturally contain underlying storylines, but also comprise fruitful visual content in the real world accompanied by rich textual descriptions, subtitles or plots, *etc*.

3.1.1 Data Source Selection

To construct the Sc2St dataset, we started by exploiting several existing movie datasets [2–8], which cover a wide range of movies with various annotations. Using existing datasets brings two advantages. First, it provides alignment with the existing dataset format. It is convenient to use a familiar dataset for research: for example, MS-COCO [39] has drawn great attention, with following work providing additional annotations [40, 41] based on the original dataset. The second advantage is the saving of significant time and labor. Some annotations, *e.g.*, subtitles and descriptions, do not change, once collected for a movie.

We carefully explored existing movie understanding datasets, such as MovieQA [3], LSMDC [6], MovieNet [7], and Condensed Movies [8]. After investigating their availability, accessibility, and richness , we eventually selected LSMDC as the data source upon which to build our dataset. LSMDC stems from an active movie understanding challenge; it has well-maintained movie videos and textural annotations. The entire LSMDC has 204 movies with various genres including action, science-fiction, family, documentary, *etc*. In particular, LSMDC provides short clips (of several seconds) with corresponding detailed descriptions. In contrast, other datasets such as [8] only contain brief descriptions for lengthy movie segments (of several minutes). Tab. 2 quantitatively compares the average clip duration, showing LSMDC has shorter clips (4.1 s). In addition, the text annotations of LSMDC were collected from two main sources, movie scripts



Script:
(1) In an office, the girl on the phone lays down the receiver (2) then crosses to the window. (3) A limb pokes through the wall. (4) Three of the limbs doing the work, the fourth holding AUNT MAY. (5) OCTAVIUS is hammering up the outside of the office building. (6) Spider man appears above him. (7) High up the building, OCTAVIUS drops AUNT MAY. (8) A web come from Spider man catches her, and AUNT MAY springs back up. (9) Below her, OCTAVIUS crashes SPIDER–MAN against the building. (10) Amazingly, SPIDER–MAN still has a battle on his hand.

Fig. 2 A storyboard example from the *Spiderman* movie. This example is composed of 10 keyframe-description pairs, but it can be used to construct storyboards of flexible length. The characters' names are shown in uppercase for clarity. Note that each description can be part of its original LSMDC clip-level description, such as descriptions (1) and (2). This enables pairing with fine-grained keyframes in our dataset.

and audio descriptions (AD). The latter are usually used to help people with visual impairments understand movies, so are more accurate and descriptive. They are well (manually) aligned with corresponding movie clips to form the video-text pairs. Unlike other annotation formats (*e.g.*, movie synopsis, subtitles, and plots), the annotated text here can provide a narrative description of the storyline as well as the key visual elements in movie clips. Finally, after excluding the movies without accessible annotations, *e.g.* blind test sets, we obtained 165 movies in total, each with hundreds of video clips and textual annotations.

Although each movie in LSMDC is split into hundreds of clips of varying duration (typically a few seconds), the video clips cannot be directly used for our task since they contain consecutive frames. Also, the corresponding text description for one clip generally contains multiple sentences and needs further splitting. We then perform several processing steps to build a keyframe-based story dataset with fine-grained textual description: (i) *text processing* which parses sub-text from the original description, (ii) *keyframe selection* that extracts the most representative keyframe image representing the given text description, and (iii) *story formation* to build stories of a specific length for the retrieval task. We next describe these three processing steps in detail.

3.1.2 Text Processing

The descriptions of the original text-based-video pairs in LSMDC usually comprise multiple sentences or sub-sentences. Directly using them would require selecting multiple keyframes covering different scenes, while we believe that using elementary sentence and image pairs is likely to be more helpful to finer text-based-image understanding. In order to construct fine-grained matching, we designed an automatic two-step text processing procedure. First, we perform language analysis using the spaCy tool to parse and split the text into sentences. Second, we further split the sentences into sub-sentences using a rule-based method. We collect a list of conjunctions (*e.g. then*), delimiters (*e.g.*, *semicolon*), and other particular symbols (*e.g.*, consecutive dashes) based on analysis of the text. An example is shown in Fig. 2, where scripts (1) and (2) originate from the same sentence, and the corresponding frames are well-matched to each sub-sentence. It clearly demonstrates that fine-grained extraction of sub-sentences assists selection of more specific keyframes. We next describe how the keyframes are selected and aligned with the given text.

3.1.3 Keyframe Selection and Alignment

Videos are composed of successive frames. We first sample raw frames at a sampling rate of 5 frames per second from the original LSMDC video clips. The sampled frames are usually redundant, and often contain similar or even duplicated visual information. To select keyframes with high expressiveness and well-matched to the text, an intuitive idea is to use rule-based methods, such as selecting a frame with a fixed position (*e.g.* the middle), or randomly selecting a frame. However, a raw clip in LSMDC may contain unrelated frames at its start or end, due to errors in labeling. For example, Fig. 3 shows that two frames of Dobby are wrongly included in the gateau scene (Harry Potter and the Chamber of Secrets). Therefore, we use a two-step process: semantic alignment based on scoring the frame-text match, and human screening to avoid errors.

We compute the text-image similarity using a universal pre-trained model [42], which has also been used for

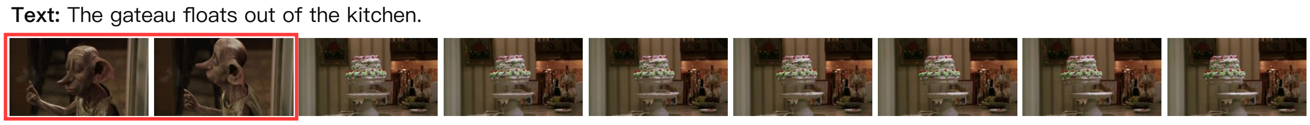


Fig. 3 In the LSMDC dataset, frames at the edge of a shot can be unrelated to the scene described by the text due to manual clip segmentation errors. Red boxes show such frames belonging to the previous shot.

constructing multi-modal datasets [43, 44] by aligning images and text. Then the frames are sorted based on their cosine similarity to the query text. By default, the top-ranked frame is taken as the keyframe, as it shares the most semantic features with the text instance.

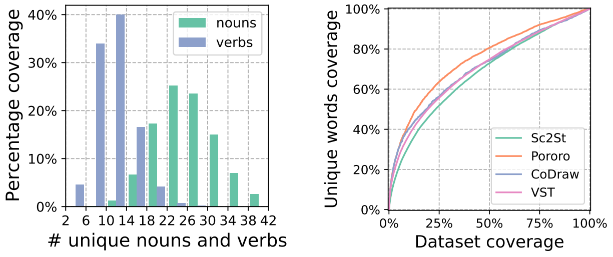
We then perform a manual screening process to check whether the selected keyframes are reasonable from a human point of view. Specifically, we show the selected keyframes based on image-text similarity to three experienced annotators and ask them to check if the keyframes (i) match the text, and (ii) are more representative than other similar frames. If all annotators agree with the initial selection, it is kept. If not, voting is used to determine the final keyframe selection. For even votes, further discussion is conducted, followed by further voting.

3.1.4 *Story Formation*

After obtaining fine-grained text descriptions and aligned keyframes, the final step is to construct the Sc2St data samples in story form. Specifically, a *storyboard* is composed of sequences of consecutive keyframes, with each keyframe paired with a sentence or sub-sentence, so that all the sentences form the script that tells the whole story. The clips from which the keyframes are derived should be temporally close ($<$ 10 seconds) to ensure semantic storyline coherence. In our implementation, we set a fixed image sequence length (10 in our experiment) for each storyboard to balance the feasibility and difficulty of the task. In detail, story formation has two main steps, clip grouping and within group story formation.

The video clips in the original LSMDC are cut from movies, and neighbouring clips may have a gap between them. Its length may be found by examining the current clip's starting timecode and the previous clip's ending timecode. A small interval indicates that these neighbouring clips very likely belong to the same scene. After experimentation, we chose 10 seconds as the threshold for grouping clips.

For each group, starting from the first clip, we add each successive clip's associated keyframes with paired sentences to form a storyboard until its length is at least the required story length. If the storyboard has a greater length, we remove the extra keyframe-text pairs at the end to provide the required storyboard length. After a storyboard is generated, we move to the next clip and restart grouping.



(a) Sc2St word distribution

(b) Diversity Comparison

Fig. 4 (a) Distribution of unique verbs and nouns within the storyboard scripts in the Sc2St dataset. (b) Percentage coverage of unique words (y-axis) by dataset coverage (x-axis), compared to other image-sequence-based story datasets. VST is the Visual Storytelling [14] dataset. Curves with higher slopes mean more unique words are used up as dataset coverage grow. The Sc2St dataset has a more gentle curve and thus has more diverse words in the text.

Here, we specify that a story has a fixed length, following [13, 14] for easier evaluation and benchmarking. Note that we could easily generate storyboards with flexible lengths for different experiments and scenarios using the above methods. Fig. 2 presents an example of a storyboard with its script consisting of sentences or sub-sentences. More storyboard samples can be found in the appendix.

3.2 Dataset Analysis

3.2.1 *Overview*

The final Sc2St dataset consists of ≈20.4k storyboards covering ≈61.2k distinct keyframe images, ≈204k sentences (≈44k distinct sentences), ≈21k unique words, and ≈2.9k characters. Taking the 10-image storyboard as an example, most scripts in Sc2St dataset contain 80 to 127 words (at the 20 and 80 percentile, respectively), with the average words around 108. Using parts-of-speech analysis, Fig. 4(a) shows the distribution of unique nouns and verbs for each script and Fig. 5 illustrates the word cloud of the most frequently used verbs, nouns, attributes, and characters. The following analyzes the dataset characteristics from various perspectives.

3.2.2 *Descriptive and Fine-Grained Text*

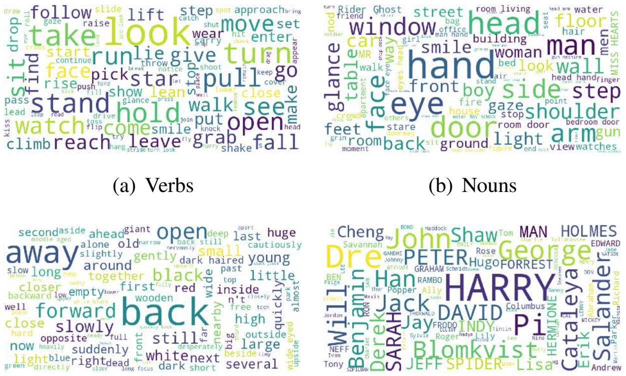
Compared to existing movie or story understanding datasets, the text annotations in our Sc2St dataset have two differentiating features: they are*descriptive* and *fine-grained*. Tab. 1 shows the text's characteristics. Most existing datasets use high-level (*e.g.* plot or synopsis) or raw (*e.g.* subtitles) textual annotations. The Sc2St dataset relies on descriptive sources (audio description and movie scripts) for descriptive and visual-content aware purposes. Although other datasets such as MovieGraphs and Condensed Movies also contain descriptions, the corresponding video lengths are much longer. We categorize the visual units into short clips, long clips, and segments based on increasing video duration; movie clips have detailed text annotation while segments have coarse annotation. Tabs. 1 and 2 summarize information of visual units. Only LSMDC and PororoQA contain short clips. Our Sc2St dataset has finer text than the original LSMDC, *e.g.*, there are 1.08 sentences per clip (#sents/unit), compared to 1.0 in LSMDC: on average a video clip is described by more (sub-)sentences. The duration per sentence (dur./sent) values also verify this. Considering image-based datasets, we have similar statistics of words and sentences to the Visual Storytelling dataset.

3.2.3 *Text Diversity*

In terms of text diversity, we compare our Sc2St dataset to other similar story datasets which contain sequential text-image pairs, including Pororo [9], CoDraw [10] and Visual Storytelling (VST) [14]. Fig. 4(b) shows the cumulative coverage of unique words (y-axis) by coverage of dataset (x-axis) on shuffled text annotations for each dataset. A steep curve at the start (left-hand x-axis) reflects lower diversity. The result shows that word distribution in our dataset is more even: at 25% coverage of dataset samples, our dataset covers only 52% words compared to Pororo (62%), CoDraw (57%), and VST (56%). Thus, the Sc2St dataset has more diverse descriptions than the others.

3.2.4 *Story Coherency*

In terms of story understanding, datasets in the movie domain unusually use videos as visual content while facing several challenges to reflect a clear story structure. First, minutes-long videos always contain many shots, resulting in complex and distant story structures with various backgrounds and characters [3, 7, 8]. Second, the discrepancy between dense video format and simple textual description leads to poor cross-modality alignment [7, 8]. Third, shorter video clip-text pairs often ignore the larger context [6]. These problems were noted in [9] and a cartoon-video-based dataset Pororo was proposed to leverage the simple storyline in cartoon arts to maintain story coherency. Instead of using videos, a derived image-version Pororo-SV adopts the images extracted from videos to build the stories for Story Visualization [13]. Similarly, others such as CoDraw and Visual Storytelling use image sequences as a way to reflect the context. Our proposed Sc2St dataset also uses a sequential approach to tell stories, while the visual movie content is richer than in cartoon datasets and more coherent than in an open domain. Statistics are shown in Tab. 1(last column).



(a) Verbs (b) Nouns

(c) Attributes (d) Characters

Fig. 5 Word clouds of the most frequent verbs, nouns, attributes, and characters in our Sc2St dataset.

4 Benchmark Evaluation

In this section, we first elaborate on the task definition with a specifically designed evaluation protocol, and then evaluate both baselines and the state-of-the-art in text-based-image and video retrieval on the proposed Sc2St dataset. We further propose our own approaches with three variants targeting the contextual retrieval task. Finally, we show how we adapt the evaluation protocol to human participants.

4.1 The Contextual Retrieval Task

4.1.1 *Task Definition*

Given a paragraph of movie script with s sentences $\mathcal{S} = (S_1, \ldots, S_s)$, the storyboard history with a series of images $\mathcal{I} = (I_1, \ldots, I_{t-1})$, and a list of 100 candidate images $\mathcal{C}_t = (C_t^{(1)}, \ldots, C_t^{(100)})$, the output should rank $\mathcal{C}_t$ for the most suitable I_t. Here t denotes time steps or rounds in the storyboard. In our setting, $2 \leqslant t \leqslant 10$. The reason we start from $t = 2$ is twofold. First, there is no previous frame at $t = 1$, which impedes models from retrieving a reasonable keyframe. Second, a frame at $t = 1$ is necessary for human evaluation at $2 \leqslant t \leqslant 10$, and thus for learned models to make comparisons. The maximum time step is 10 because each storyboard contains 10 keyframes, meaning there are 9 rounds of retrieval for every storyboard.

4.1.2 Dataset Splits

The Sc2St dataset contains 20413 storyboards, each of which has 10 images and needs 9 rounds of experiments. The dataset is split into 16330 for training (80%), 1022 for validation (5%), and 3061 for testing (15%). Thus, there are 183717 training, 9198 validation, and 27549 testing rounds in total.

4.1.3 Candidate Keyframes

We prepare the candidates as follows. For each ground truth image in a storyboard sample, the candidate set includes 1 correct image and 99 incorrect images of two kinds: *Similar images* are about ≈70% of the total. We first extract the 1024-dimensional features for all keyframe images using the 121-layer DenseNet [45], then compute the cosine similarity matrix over all keyframes. Then, each keyframe is assigned a set of most similar candidates, which are chosen from the same movie and other movies, with about 30% from the same movie to ensure a certain level of difficulty. *Random images* make up the remaining ≈30%, and are randomly selected from the other keyframes from the current movie (series) and other movies, again in the same ratio.

Note that there are no overlapping candidates across the different data splits to avoid data leakage. The candidates can include those that do not belong to any storyboard, i.e. isolated ones not qualified to form a story.

4.1.4 Evaluation Metrics

Each round of retrieval is prepared with a list of 100 keyframe candidates. The tasks are automatically evaluated using retrieval ranking scores on the candidate lists, which include (i) recall@k, the recall (percentage) of the top k ranked (higher is better), (ii) mean rank (lower is better), and (iii) mean reciprocal rank (MRR) (higher is better).

4.2 Baselines

We consider two baselines to evaluate whether methods are better than chance. The *prior* baseline is given by random results over the candidates without using any inputs, and the *similarity* baseline comes from results obtained by descending cosine similarity scores between candidate images and the image in the last round.

4.3 State-of-the-Art

No methods directly target our task, and the most closely related research targets text-based-image and video retrieval. Considering the types of methods and involvement of context encoding, we classify existing methods into four groups: text-based-image retrieval , video retrieval, pre-training-based visual transformers, and contextual retrieval, respectively.

*4.3.1 Text-based-image Retrieval (*I-*)*

We compare to two recent text-based-image retrieval methods: SCO [46], that learns sentence-image similarity, and CAMP [19], for word and region-level similarity learning. For fairness of comparison, we use Faster R-CNN [47] to extract region-level visual features and [48] for encoding word embeddings.

*4.3.2 text-based-video Retrieval (*V-*)*

Mixture of embedding experts (MoEE) models are widely used in text-based-video retrieval [23, 49]. A standard procedure is to use a weighted combination of multiple expert embeddings for video representations to learn text-video similarity. Although our data modality is images, we treat it as a single-frame video following [26]. We use the following experts for keyframe representation: scene features using the DenseNet161 model [45] pre-trained on the Places365 dataset, object features using the SENet154 [50] model pre-trained on ImageNet, and a character embedding that encodes the top-100 characters mentioned in the text.

*4.3.3 Pre-training-based ViT (*P-*)*

Recently, vision-language transformers have been widely used in multi-modal alignment based on pre-training. We choose to compare to the representative UNITER [21] and OSCAR [22] models. Both adopt object tags and regions detected in images to better learn the text-image alignment. We first extract the detected objects with region features from the keyframes using Faster R-CNN [47]. Then, we fine-tune the pre-trained models by feeding them with the text-keyframe pairs and object features. The [*CLS*] token is used as input for the following retrieval task.

*4.3.4 Contextual Retrieval (*C-*)*

We further compare to methods involving contextual retrieval, which exist for text-based-image or text-based-video retrieval. For the former, we compare to a neural story illustration method [11], StoryShow. It uses a hierarchical GRU network to learn a representation for the input story while keeping coherence between sentences to retrieve a sequence of ordered images, which is similar to the setting for the Sc2St task. For the text-based-video domain, we compare to the Contextual MoEE (CMoEE) [8], which learns the similarity score between text and video using weighed expert features from the current and past video clips. We replace the original video clips with keyframes and use the same experts explained in the MoEE model. As our contextual retrieval task involves

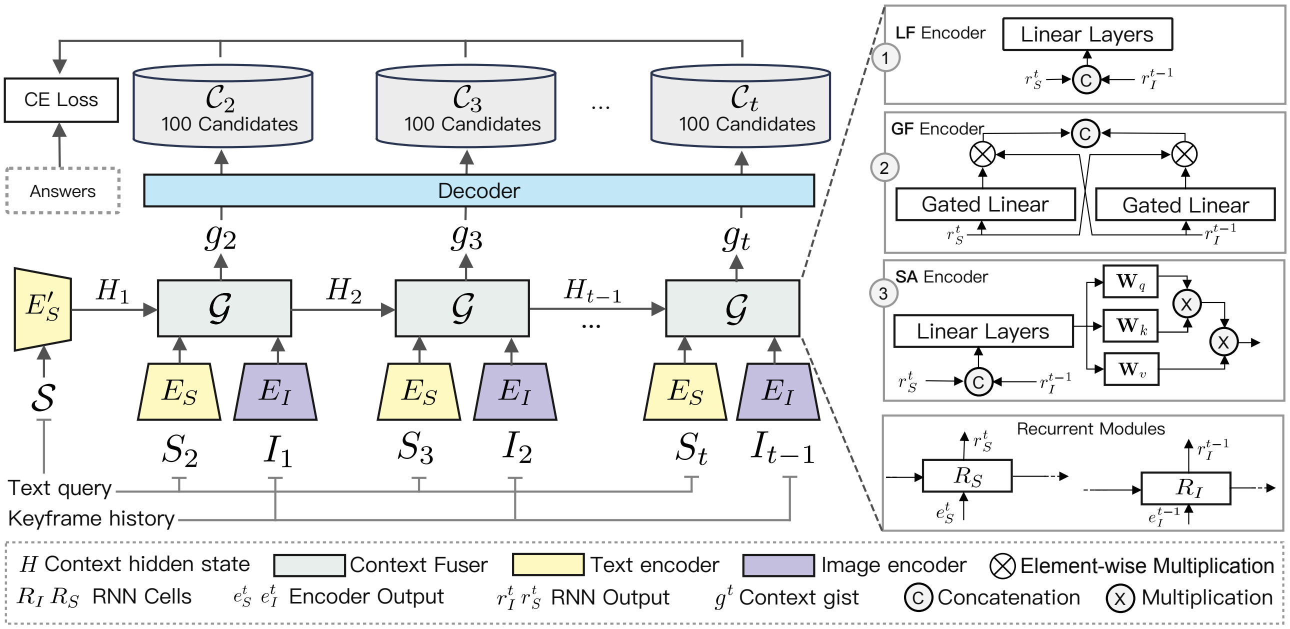


Fig. 6 Proposed recurrent model architecture using three contextual fusion encoders. After encoding the text query and keyframe separately, a context encoder $\mathcal{G}$ is used to capture the context gist g from the recurrent outputs (r_S or r_I) at each time step t. The two recurrent modules are implemented using LSTM cells to encode both streams. Then a fusion encoder (*LF*, *GF*, or *SA*) takes the two recurrent outputs and fuses them in different ways, as follows: *LF* performs late fusion using concatenation, *GF* uses two gates for cross-gating the information from the two modalities, and *SA* uses self-attention on the concatenated recurrent outputs. At each time step, there is a prepared keyframe candidate list C_t used for optimizing the model using the cross-entropy (CE) loss.

text and frame context, these two methods have a better fit to our task than other methods.

4.4 Proposed Methods

4.4.1 Approach

The storyboards contain rich visual and textual context, namely the keyframes and text descriptions. To better capture the contextual information from both sequences, we base our model on a recurrent architecture. Fig. 6 illustrates the framework. Specifically, at each time step t, the text query S_t and last-round image I_{t-1} (the keyframe history) are separately encoded using a text encoder E_S and image encoder E_I. Then, a context encoder $\mathcal{G}$ is utilized to capture the contextual gist g_t by tracking the hidden states from both image and text streams. In the first round, a script encoder E'_S encodes the entire movie script to form the initial hidden state H_1. The following decoder takes the gist as input and finally computes the dot product similarity with the given candidates.

In the implementation, the image encoder is derived from a pre-trained Inception-v3 [51] model. It serves as a general feature extractor that converts images in the storyboard history $\mathcal{I} = (I_1, \ldots, I_{t-1})$ into two types of features: a last-layer feature map $\mathcal{F} = (f_1, \ldots, f_{t-1})$, and its global feature vector $\bar{\mathcal{F}} = (\bar{f}_1, \ldots, \bar{f}_{t-1})$ by applying global pooling to $\mathcal{F}$. The global image features $\bar{\mathcal{F}}$ are exploited as initial keyframe features. Note that we append additional layers to the image encoders to fine-tune them. The text encoder is a pre-trained uncased BERT model [48], and we use the pooled features for text representation. $\mathcal{G}$ is the core part of our method, and has two parts: recurrent modules and fusion modules. The recurrent modules use LSTMs to encode sequential text and keyframe features separately, while the fusion modules aim to fuse the recurrent outputs to generate the context gist. We use three mechanisms to fuse the recurrent outputs, as follows.

4.4.2 Late Fusion (LF)

In the late fusion encoder, the text and image features are directly concatenated and then processed by a multilayer perceptron (MLP). This simple approach fuses the two modality features into a joint semantic space (g).

4.4.3 Gated Fusion (GF)

The gated fusion encoder has a gating mechanism that controls what information is passed on or forgotten, as in a gated linear layer (GLU) [52]. Here we propose a cross-gating mechanism: using two gated layers for filtering image history information by text information and filtering text information by image history information. The two outputs are then concatenated as the context gist, as formulated by:

$$g_t = G_S\left(r_S^t\right) r_I^{t-1} + G_I\left(r_I^{t-1}\right) r_S^t \tag{1}$$

where G is the gated module including linear transformation and a sigmoid layer, r_S^t and r_I^{t-1} are recurrent outputs for text at current time step t and keyframe history at the previous time step $t-1$.

4.4.4 *Self-Attention Fusion (SA)*

We leverage an attention mechanism [20] to connect the visual and textual information. Specifically, self-attention (SA) is used to bridge the two recurrent outputs where we adapt the non-local concept from [53] to implement the SA module for our data inputs. Firstly, image and text contexts are concatenated and fed into an MLP to get the N-channel unified representations $\mathbf{u} \in \mathbb{R}^{U\times N}$, where U is the unified feature dimension. Then, attention is realized by introducing three linear transformations: $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$. The self-attention computation on each channel of $\mathbf{u}$ is formulated as:

$$\mathbf{u}_j = \mathbf{W}_v \sum_{i=1}^{N} w_{j,i}\mathbf{u}_i, \tag{2}$$

$$w_{j,i} = \frac{\exp\left(\alpha_{i,j}\right)}{\sum_{k=1}^{N}\exp\left(\alpha_{i,k}\right)}, \tag{3}$$

where $\alpha_{i,j} = \left(\mathbf{W}_q\mathbf{u}_i\right)^T\left(\mathbf{W}_k\mathbf{u}_j\right)$. Based on the obtained context gist g_t, the decoder first computes the dot product similarity to the 100 keyframe candidates' features and uses a softmax layer to get the classification score (posterior probabilities) over all candidates. Cross-entropy loss is then adopted for optimization.

4.5 Human Evaluation

To evaluate human performance on the contextual retrieval task, we built an online user study system (Online Human Evaluation: https://sc2st.com) using the same testing set. As the testing set includes ≈3k storyboard samples covering ≈27k rounds of experiments, we randomly chose 30 samples from it to reduce the testing size, resulting in a total of 270 rounds of retrieval. To align the human evaluation with the ranking-based evaluation protocol as well as to make it practical to conduct, for each retrieval round, we allow participants to choose at least 1 but up to 10 images instead of exactly 10 images according to their confidence. In this way, it is convenient and efficient when participants are more confident about the already chosen and ranked images, since they do not need to rank more to make up 10 images, and the rest are automatically filled by random frames. Overall, we obtained human evaluation results from 14 participants on 321 data samples.

Table 3 Retrieval results for baselines, state-of-the-art methods, our methods, and human performance. ↑ (↓) means a higher (lower) value is better.

Method	R1↑	R5↑	R10↑	Mean↓	MRR↑
Prior	0.51	3.30	8.01	49.24	0.043
Similarity	7.72	16.69	23.55	43.10	0.137
I-SCO	3.89	18.6	33.27	24.72	0.1304
I-CAMP	4.36	20.57	35.53	23.66	0.1425
P-UNITER	9.0	31.36	49.72	16.57	0.203
P-Oscar	11.48	35.72	54.28	15.12	0.240
V-MoEE	5.21	21.05	35.0	24.29	0.149
C-StoryShow	10.05	27.17	40.95	21.06	0.201
C-CMoEE	12.92	38.54	58.55	15.84	0.234
Ours-LF	29.0	49.64	61.06	14.88	0.395
Ours-GF	29.02	*49.75*	60.94	*14.30*	0.396
Ours-SA	*30.22*	49.55	*61.0*	14.31	*0.398*
Human	38.01	57.6	73.58	-	-

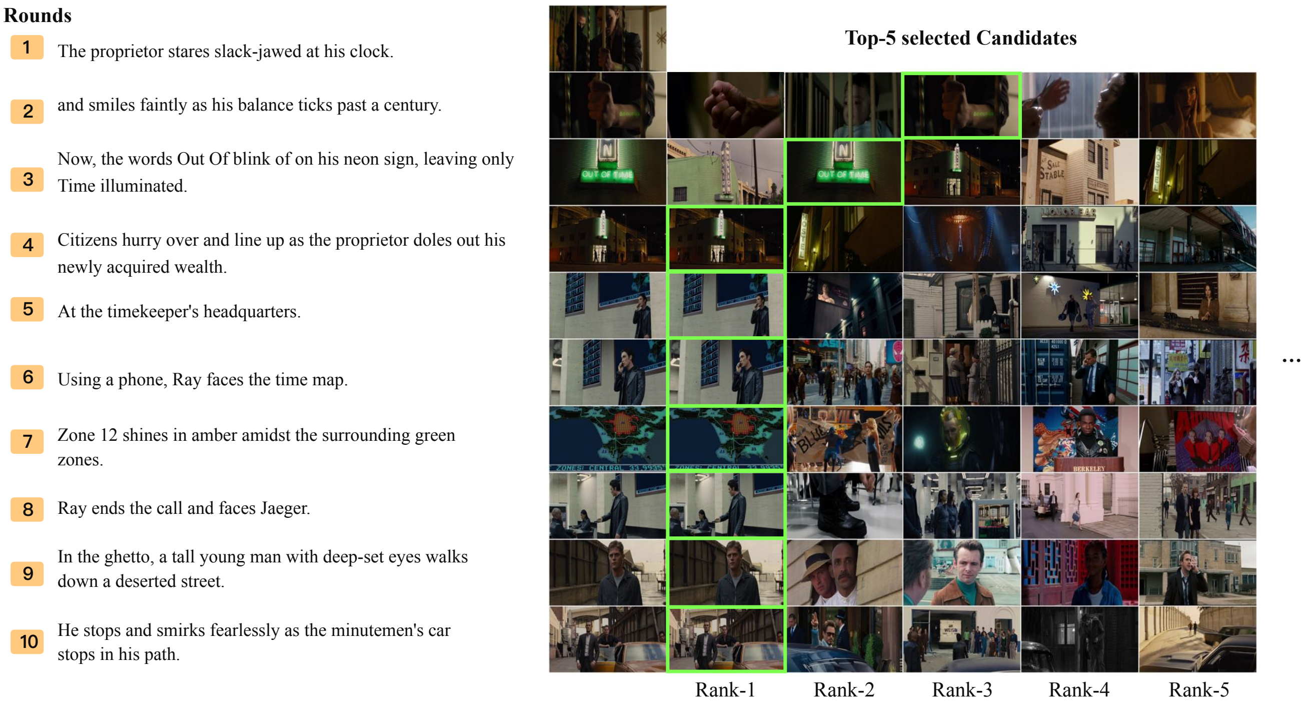
5 Results and Discussion

5.1 Quantitative Results

Tab. 3 summarises quantitative results of evaluating the various methods. Our methods perform favorably against all other existing methods under all metrics. There are subtle performance differences between the three context encoders in terms of R5 and R10, while the GF and SA encoders perform slightly better than the LF encoder for R1. The SA encoder outperforms both the LF and GF encoders under R1. The pre-training-based transformers (UNITER [21] and Oscar [22]) display superior performance to the classic similarity-based text-based visual retrieval methods (SCO [46], CAMP [19] and MoEE [49]). For context-based methods (C-StoryShow [11] and C-CMoEE [8]), the better performance of C-CMoEE indicates the visual context has a greater weight than textual context; we further validate this in the ablation study (in Section 5.3). Human subjects achieve leading results in all metrics. Mean and MRR results are absent for the user study, as these metrics need rankings over all the candidates which is unachievable for the user study.

5.2 Qualitative Results

Fig. 7 shows the top-5 selection results over an entire story example by our model using the LF context encoder. Given the initial keyframe with its descriptions, for each round, the model needs to predict the possible keyframe conditioned on the text description and the frame history. It can be seen that from the top selected frames, the visual features share a semantic similarity. For example, the candidates in the third row concern a *neon sign* while Rank-1 and Rank-4 in the 6th row show a scene of *a person using a phone*, *etc*. Only using text-based-image similarity is insufficient, as there are



Rounds

1 The proprietor stares slack-jawed at his clock.

2 and smiles faintly as his balance ticks past a century.

3 Now, the words Out Of blink of on his neon sign, leaving only Time illuminated.

4 Citizens hurry over and line up as the proprietor doles out his newly acquired wealth.

5 At the timekeeper's headquarters.

6 Using a phone, Ray faces the time map.

7 Zone 12 shines in amber amidst the surrounding green zones.

8 Ray ends the call and faces Jaeger.

9 In the ghetto, a tall young man with deep-set eyes walks down a deserted street.

10 He stops and smirks fearlessly as the minutemen's car stops in his path.

Fig. 7 Qualitative retrieval results. The top-5 selected keyframes are shown given the script, denoted by green boxes.

similar candidates making the task challenging, while our context-aware method can leverage the visual and textual history context for better retrieval. Later rounds receive more contextual information and the results tend to be better (rounds 4–10). We now quantitatively verify this observation.

5.3 Effectiveness of Context Encoding

The core part of our proposed approaches is the recurrent architecture using context encoders to capture the contextual information. In order to better compare whether the method can effectively use the context, Tab. 4 presents the results for early (rounds 2–4), middle (rounds 5–7), and late (rounds 7–9) temporal stages in storyboards using our method (LF). In comparison, the MoEE results are also shown for each temporal stage. Results demonstrate that our method performs better in middle or late temporal stages than in earlier stages, meaning that availability of more history information for later rounds is important, and our model can successfully utilize previous context for the retrieval task. However, non-contextual models (like MoEE) do not show this change, and indeed the performance in earlier stages is better than in later stages.

Table 4 Results for different temporal stages using our method with LF context encoder. Later temporal stages have better results than earlier stages, indicating the historical context is significant in retrieval.

Rounds	R1↑	R5↑	R10↑	Mean↓	MRR↑
MoEE-Early	5.18	20.08	35.14	24.32	0.150
MoEE-Middle	4.89	20.12	34.65	24.43	0.144
MoEE-Late	5.21	20.09	35.0	24.31	0.147
SA-Early	28.27	47.97	59.83	14.81	0.386
SA-Middle	31.18	50.56	62.03	13.96	0.412
SA-Late	29.24	48.33	61.12	14.05	0.408

Table 5 Results of using textual- context only (-T), visual-context only (-V) and both (-Full), in the LF model.

Method	R1↑	R5↑	R10↑	Mean↓	MRR↑
LF-T	10.82	27.54	40.93	21.0	0.2066
LF-V	19.01	44.63	60.14	16.91	0.2993
LF-Full	29.0	49.64	61.06	14.88	0.395

5.4 Effectiveness of Dual Context

To show the contribution of visual and textual context used in our methods separately, we designed experiments that use only the visual context and only the textual context, and compare them to the full model using both contexts. The results are shown in Tab. 5 using the LF fusion module. They show that the usage of visual context, namely the frame history, can alone perform better than the textual context for all metrics, suggesting that visual elements of the context are more important. The performance is further boosted by adding the textual context, especially in R1, by about 52%.

1. HARRY steps through the doorway and into a strange new world. 2. Diagon Alley is a narrow crooked lane hustling with witches and wizards in flowing cloaks and pointed hats. 3. HAGRID towers over them all. 4. They pass tiny shops packed full of fascinating and unusual merchandize. 5. A selection of wizard's cloaks and hats. 6. A pet shop full of owls and cages. 7. An owl sits on a purge outside the shop. 8. A bat stretches its wings above HARRY's head. 9. A group of children gaze enviously at a sleek broomstick in a shop window. 10. HARRY looks worried.

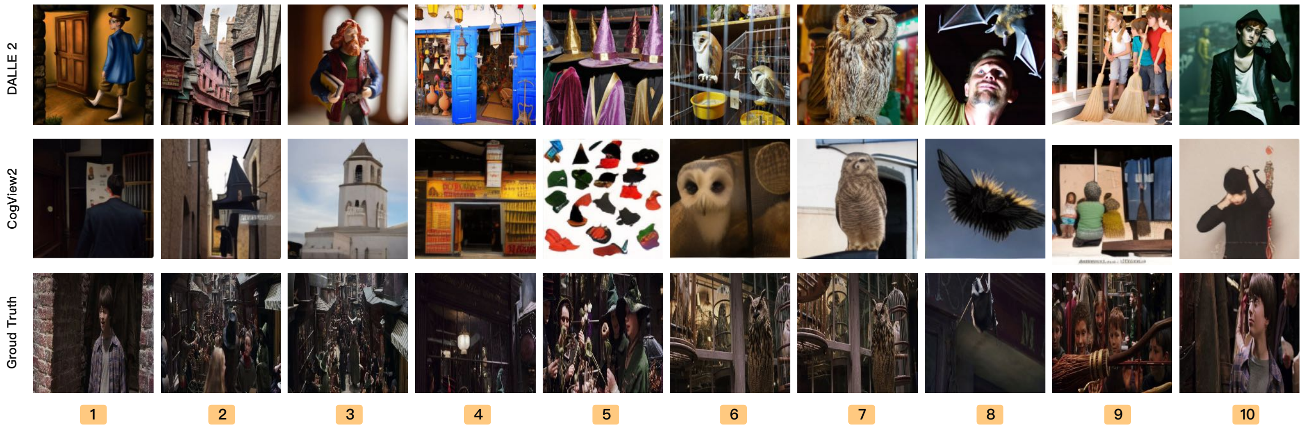


Fig. 8 Reviewing the text-to-image generation results on our dataset using Dall·E-2 and CogView2.

5.5 Exploring Potential Applications

As the Sc2St dataset has a clear story structure, we hope to further explore its application to other movie-related scenarios besides contextual retrieval. One possible application may be script-guided storyboard generation, unlike the usual text-to-image (T2I) task, where a sequence of images needs to be generated. Currently, T2I is still a challenging task since most research has focused on single object (e.g. birds or flowers) generation from text, and the quality of generated complex scenes is not ideal [32, 54]. This is more challenging for the movie domain which involves various characters, objects, and scenes. Recent advances in T2I are driven by scaling models on large datasets. These models, with billions of parameters, and trained from abundant data using hundreds or thousands of GPUs, show the potential to generate realistic images from text [33–35, 55]. We thus examined the generation quality of recently available state-of-the-art T2I models: Dall·E-2 [35] and CogView-2 [34]. Specifically, each keyframe in a storyboard is generated one by one given its paired text. Some results are shown in Fig 8. Note that the generated images are carefully picked since each sentence can generate multiple images from different random initialization, and we manually screened the most suitable images. Dall·E-2 can generate more realistic images than CogView-2. Simple scenes (*e.g.* the owl) can be well synthesized by both models while complex scenes are harder for CogView-2. The generated images lack coherence, having inconsistent styles, which is to be expected since no contextual information is used. With T2I benefiting from large-scale modeling, we thus hope our proposed dataset can be applied to fill the gap for more coherent and realistic storyboard generation in future.

5.6 Limitations and Future Work

First, the movies included are limited. Our Sc2St dataset can provide fine-grained storyboards based on 165 movies. Inspired by the recent large-scale movie dataset [7], there is potential to include more movies with our storyboard annotations. Second, the evaluation of contextual generation can be further investigated. We carefully designed the benchmark evaluation for the Sc2St contextual retrieval task. However, automatic perceptual evaluation of the generated results remains challenging, and we leave it for future work. Third, although we discussed potential text-to-image generation and qualitative results using our dataset, other applications such as storytelling, and video creation/editing applications would also benefit from our dataset.

6 Conclusions

In this paper, we proposed a new script-to-storyboard dataset (Sc2St) together with a contextual retrieval task in the movie domain. The new dataset features a new data form called a storyboard, which consists of sequential keyframe images with corresponding textual descriptions. A storyboard has the advantage of an explicit, clear, and coherent story structure over the implicit storyline in movies. Compared to existing movie datasets, the Sc2St dataset contain fine-grained, highly diverse text annotations. The newly annotated keyframes are semantically matched to the text. Using the new dataset, we have benchmarked the contextual retrieval task with an automatic ranking-based evaluation protocol. We have proposed baselines with three variants to accomplish the task and compare them to state-of-the-art methods as well as human performance. Quantitative results demonstrated that our approach performs better by successfully leveraging contextual information from both the text and image history. Finally, we explored the potential of the generation task using our dataset.

A.1 Implementation

Our proposed models were implemented using the deep learning framework PyTorch [56]. Adam [57] was used as optimisation method, with learning rate set to 0.0003 and scheduled with a cosine annealing strategy. For all models, the batch size was set to 128. To select a best model, mean ranking performance evaluated on a hold-out validation set was used.

A.2 Additional Storyboard Samples

Figs. A1–A3 show more storyboard samples from different movies, demonstrating that the keyframes in the storyboards summarize the condensed visual information reflected in the textual description. The sequences of keyframes provide a coherent story, which one can understand without the original long and redundant video information.

A.3 Human Evaluation Interface

The human evaluation interface is shown in Fig. A4. A demonstration is provided at https://sc2st.com.

Acknowledgements

This research was supported by RCUK grant CAMERA (EP/M023281/1, EP/T022523/1), the Centre for Augmented Reasoning (CAR) at the Australian Institute for Machine Learning, and a gift from Adobe.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval

This study does not contain any studies with human or animal subjects performed by any of the authors.

References

[1] Huang Q, Xiong Y, Xiong Y, Zhang Y, Lin D. From trailers to storylines: An efficient way to learn from movies. *arXiv preprint arXiv:1806.05341*, 2018.

[2] Gu C, Sun C, Ross DA, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, et al.. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 6047–6056.

[3] Tapaswi M, Zhu Y, Stiefelhagen R, Torralba A, Urtasun R, Fidler S. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 4631–4640.

[4] Vicol P, Tapaswi M, Castrejon L, Fidler S. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 8581–8590.

[5] Xiong Y, Huang Q, Guo L, Zhou H, Zhou B, Lin D. A graph-based framework to bridge movies and synopses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 4592–4601.

[6] Rohrbach A, Torabi A, Rohrbach M, Tandon N, Pal C, Larochelle H, Courville A, Schiele B. Movie description. *International Journal of Computer Vision*, 2017, 123(1): 94–120.

[7] Huang Q, Xiong Y, Rao A, Wang J, Lin D. MovieNet: A Holistic Dataset for Movie Understanding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349, 2020, 709–727.

[8] Bain M, Nagrani A, Brown A, Zisserman A. Condensed Movies: Story Based Retrieval with Contextual Embeddings. In *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part V*, volume 12626, 2020, 460–479.

[9] Kim K, Nan C, Heo M, Choi S, Zhang B. PororoQA: Cartoon video series dataset for story understanding. In *Proceedings of NIPS 2016 Workshop on Large Scale Computer Vision System*, volume 15, 2016, –.

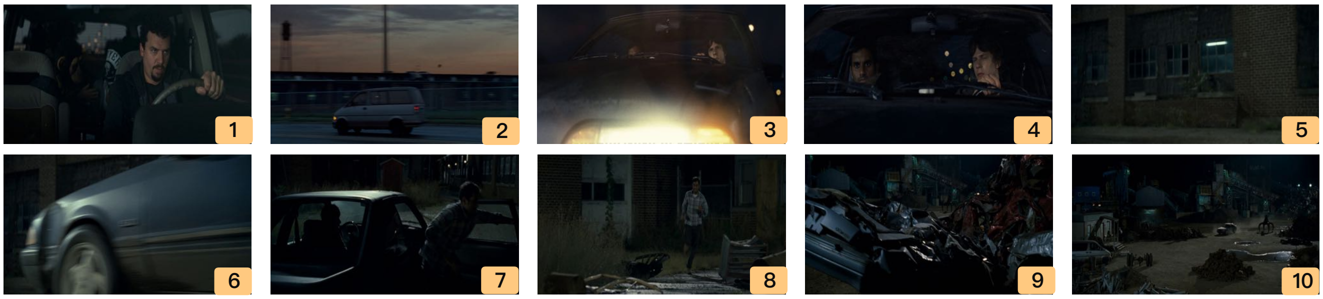
[10] Kim JH, Kitaev N, Chen X, Rohrbach M, Zhang BT, Tian Y, Batra D, Parikh D. CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, 6495–6513.

[11] Ravi H, Wang L, Muniz C, Sigal L, Metaxas D, Kapadia M. Show me a story: Towards coherent neural story illustration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 7613–7621.

[12] Chen S, Liu B, Fu J, Song R, Jin Q, Lin P, Qi X, Wang C, Zhou J. Neural storyboard artist: Visualizing stories with coherent image sequences. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, 2236–2244.

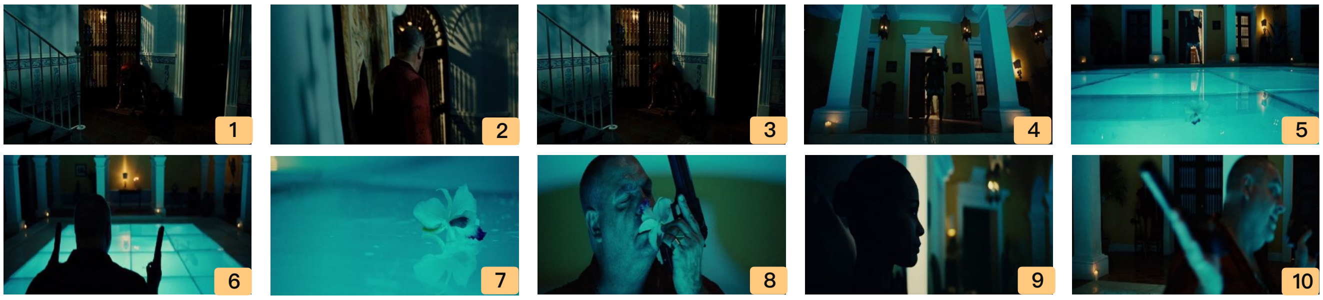
[13] Li Y, Gan Z, Shen Y, Liu J, Cheng Y, Wu Y, Carin L, Carlson D, Gao J. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, 6329–6338.

[14] Huang TH, Ferraro F, Mostafazadeh N, Misra I, Agrawal A, Devlin J, Girshick R, He X, Kohli P, Batra D, et al.. Visual storytelling. In *Proceedings of the 2016 Conference of the*



Movie: 30 MINUTES OR LESS
Script:
1. Travis repeatedly pokes at Kate's face, which is covered in a Slayer T–shirt.
2. The van speeds down an industrial road at dusk.
3. Later that night our view rises from one of the Mustang's headlights to Nick.
4. who glances determinedly at his best friend beside him.
5. Chet glances back.
6. The Mustang parks.
7. Chet steps out and the Mustang continues on down the dirt road.
8. Chet runs down a path strewn with garbage.
9. Now our view rises from a heap of crushed cars to reveal the Mustang arriving in the scrapyard.
10. Its headlights go dim and Nick steps out.

Fig. A1 Storyboard example - *30 Minutes or Less*



Movie: COLOMBIANA
Script:
1. William creeps down a flight of stairs
2. then presses his back to a wall as William reaches the bottom.
3. Bending over another dead bodyguard, William takes his gun.
4. Armed with a pistol in each hand, William steps into the courtyard.
5. William notices a flower resting three panels away on the pool's glass.
6. His nervous eyes scan the area as William walks across the pool.
7. Stopping in front of the cataleya orchid, William picks it up and squints as William examines it.
8. William lifts the flower to his nose, closes his eyes, and takes a sniff.
9. Opening his eyes, William smirks as William lifts his face from the flower.
10. Cataleya rises from a chair in the portico behind him.

Fig. A2 Storyboard example - *Colombiana*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, 1233–1239.

[15] Faghri F, Fleet DJ, Kiros JR, Fidler S. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, 2018, 12.

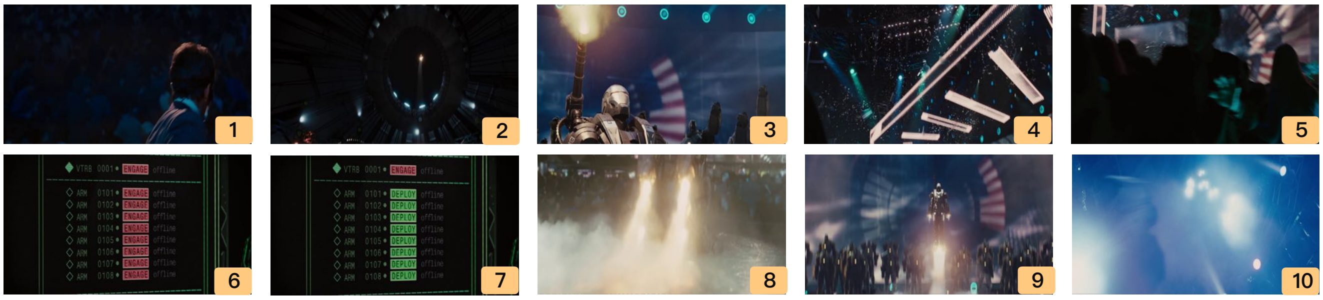
[16] Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen Y. Dual-path Convolutional Image-Text Embeddings with Instance Loss. *ACM Trans. Multim. Comput. Commun. Appl.*, 2020, 16(2): 51:1–51:23, doi:10.1145/3383184.

[17] Zhang Y, Lu H. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 2018, 686–701.

[18] Lee KH, Chen X, Hua G, Hu H, He X. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, 201–216.

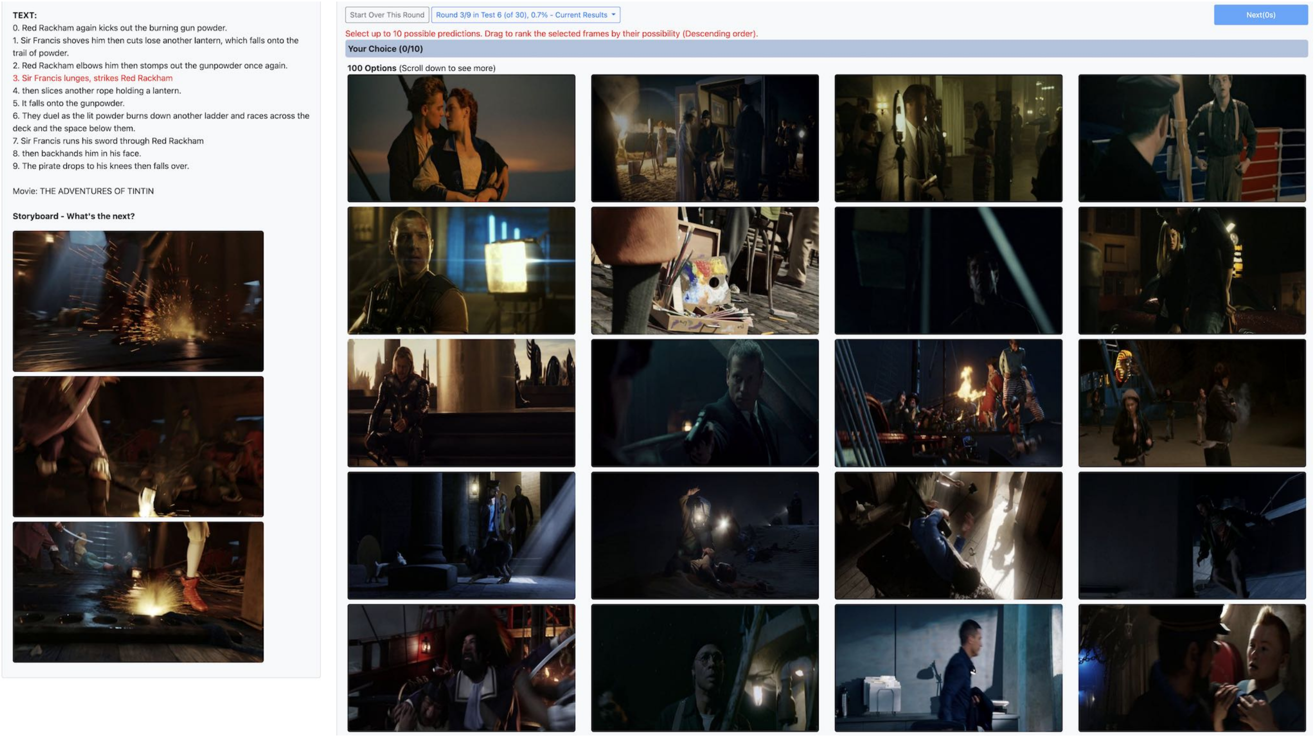
[19] Wang Z, Liu X, Li H, Sheng L, Yan J, Wang X, Shao J. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 5764–5773.

[20] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. *Advances*



Movie: IRON MAN2
Script:
1. Justin runs offstage.
2. Tony flies out through the opening in the glass roof.
3. Rhodey and the drones fire their guns at Tony.
4. The glass panels in the roof shatter.
5. showering down on the audience as people run off.
6. Ivan types on his computer.
7. Commands pop up on his computer screen: "Deploy, deploy, deploy. ".
8. Onstage, SOMEONE's thrusters fire up.
9. Rhodey rockets straight upward and out of the arena.
10. The Air Force drones fly through the roof, breaking more glass.

Fig. A3 Storyboard example - *Iron Man 2*



TEXT:
0. Red Rackham again kicks out the burning gun powder.
1. Sir Francis shoves him then cuts lose another lantern, which falls onto the trail of powder.
2. Red Rackham elbows him then stomps out the gunpowder once again.
3. Sir Francis lunges, strikes Red Rackham
4. then slices another rope holding a lantern.
5. It falls onto the gunpowder.
6. They duel as the lit powder burns down another ladder and races across the deck and the space below them.
7. Sir Francis runs his sword through Red Rackham
8. then backhands him in his face.
9. The pirate drops to his knees then falls over.

Movie: THE ADVENTURES OF TINTIN

Storyboard - What's the next?

Select up to 10 possible predictions. Drag to rank the selected frames by their possibility (Descending order).

Your Choice (0/10)

100 Options (Scroll down to see more)

Fig. A4 Human online evaluation interface.

in neural information processing systems, 2017, 30.

[21] Chen Y, Li L, Yu L, Kholy AE, Ahmed F, Gan Z, Cheng Y, Liu J. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375, 2020, 104–120.

[22] Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, et al.. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 2020, 121–137.

[23] Liu Y, Albanie S, Nagrani A, Zisserman A. Use What You Have: Video retrieval using representations from collaborative experts. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019, 279.

[24] Gabeur V, Sun C, Alahari K, Schmid C. Multi-modal transformer for video retrieval. In *European Conference on Com-*

puter Vision, 2020, 214–229.

[25] Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 1994, 6(2): 181–214.

[26] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, 1725–1732.

[27] Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, 2016, 1060–1069.

[28] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017, 5907–5915.

[29] Zhu M, Pan P, Chen W, Yang Y. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, 5802–5810.

[30] Li B, Qi X, Lukasiewicz T, Torr P. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*, 2019, 2065–2075.

[31] Liang J, Pei W, Lu F. CPGAN: Content-Parsing Generative Adversarial Networks for Text-to-Image Synthesis. In *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12349 of *Lecture Notes in Computer Science*, 2020, 491–508.

[32] Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 1316–1324.

[33] Ding M, Yang Z, Hong W, Zheng W, Zhou C, Yin D, Lin J, Zou X, Shao Z, Yang H, et al.. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 2021, 34: 19822–19835.

[34] Ding M, Zheng W, Hong W, Tang J. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers. *arXiv preprint arXiv:2204.14217*, 2022.

[35] Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[36] El-Nouby A, Sharma S, Schulz H, Hjelm D, Asri LE, Kahou SE, Bengio Y, Taylor GW. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019, 10304–10312.

[37] Xue Y, Guo YC, Zhang H, Xu T, Zhang SH, Huang X. Deep image synthesis from intuitive user input: A review and perspectives. *Computational Visual Media*, 2022, 8(1): 3–31.

[38] Wang M, Yang GW, Hu SM, Yau ST, Shamir A, et al.. Write-a-video: computational video montage from themed text. *ACM Trans. Graph.*, 2019, 38(6): 177–1.

[39] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014, 740–755.

[40] Caesar H, Uijlings J, Ferrari V. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 1209–1218.

[41] Jin S, Xu L, Xu J, Wang C, Liu W, Qian C, Ouyang W, Luo P. Whole-Body Human Pose Estimation in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12354, 2020, 196–214.

[42] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al.. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[43] Schuhmann C, Vencu R, Beaumont R, Kaczmarczyk R, Mullis C, Katta A, Coombes T, Jitsev J, Komatsuzaki A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[44] Gu J, Meng X, Lu G, Hou L, Niu M, Xu H, Liang X, Zhang W, Jiang X, Xu C. Wukong: 100 Million Large-scale Chinese Cross-modal Pre-training Dataset and A Foundation Framework. *arXiv preprint arXiv:2202.06767*, 2022.

[45] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 4700–4708.

[46] Huang Y, Wu Q, Song C, Wang L. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 6163–6171.

[47] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 2015, 28.

[48] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, 4171–4186, doi:10.18653/v1/n19-1423.

[49] Miech A, Laptev I, Sivic J. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[50] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 7132–7141.

[51] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 2818–2826.

[52] Dauphin YN, Fan A, Auli M, Grangier D. Language modeling with gated convolutional networks. In *International conference on machine learning*, 2017, 933–941.

[53] Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 7794–7803.

[54] Zhang H, Koh JY, Baldridge J, Lee H, Yang Y. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, 833–842.

[55] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 2021, 8821–8831.

[56] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al.. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 2019, 32.

[57] Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Author biography

Xi Tian is a third-year Ph.D. student at the University of Bath, United Kingdom. He obtained his master's degree in 2016 in Computer Science from the University of Bristol, UK. He received his bachelor's degree in 2014, in Computer Science from the Beijing University of Posts and Telecommunication, China. He research combines vision and language, especially focusing on sequential or contextual problems.

Yong-Liang Yang is a Senior Lecturer in the Department of Computer Science at the University of Bath. He received B.S. and Ph.D. degrees in Computer Science from Tsinghua University, Beijing. His research works are broadly in visual computing and interactive techniques, resulting in publications in top venues such as SIGGRAPH, SIGGRAPH Asia, CVPR, ICCV, CHI. He has served on program committees of multiple international graphics conferences including the Symposium on Geometry Processing, Pacific Graphics, and Solid & Physical Modelling.

Qi Wu is a Senior Lecturer in the University of Adelaide and an Associate Investigator in the Australia Centre for Robotic Vision (ACRV). He was an ARC Discovery Early Career Researcher Award (DECRA) Fellow from 2019–2021. He obtained his Ph.D. in 2015 and M.Sc. in 2011, in Computer Science from the University of Bath. His educational background is primarily in computer science and mathematics. He works on vision and language problems, including image captioning, visual question answering, visual dialog etc. His work has been published in prestigious journals and conferences such as TPAMI, CVPR, ICCV, AAAI and ECCV.