# OpenStory: A Large-Scale Open-Domain Dataset for Subject-Driven Visual Storytelling

Zilyu Ye[1]* Jinxiu Liu[1]* JinJin Cao[2] Zhiyang Chen[2,4]
Ziwei Xuan[3] Mingyuan Zhou[3] Qi Liu[1] Guo-Jun Qi[2]
[1]School of Future Technology, South China University of Technology
[2]Westlake University [3]OPPO US Research Center
[4]Foundation Model Research Center, CASIA
{zilyuye,jinxiuliu0628}@foxmail.com, caojinjin@westlake.edu.cn, zhiyang.chen@nlpr.ia.ac.cn
ziwei.xuan@innopeaktech.com, mingyuanzhou5@gmail.com, drliuqi@scut.edu.cn, guojunq@gmail.com

## Abstract

*Recently, the advancement and evolution of generative AI have been highly compelling. In this paper, we present **OpenStory**, a large-scale dataset tailored for training subject-focused story visualization models to generate coherent and contextually relevant visual narratives. Addressing the challenges of maintaining subject continuity across frames and capturing compelling narratives, We propose an innovative pipeline that automates the extraction of keyframes from open-domain videos. It ingeniously employs vision-language models to generate descriptive captions, which are then refined by a large language model to ensure narrative flow and coherence. Furthermore, advanced subject masking techniques are applied to isolate and segment the primary subjects. Derived from diverse video sources, including YouTube and existing datasets, OpenStory offers a comprehensive open-domain resource, surpassing prior datasets confined to specific scenarios. With automated captioning instead of manual annotation, high-resolution imagery optimized for subject count per frame, and extensive frame sequences ensuring consistent subjects for temporal modeling, OpenStory establishes itself as an invaluable benchmark. It facilitates advancements in subject-focused story visualization, enabling the training of models capable of comprehending and generating intricate multimodal narratives from extensive visual and textual inputs.*

## 1. Introduction

The domain of artificial intelligence has been predominantly captivated by the advent of awesome generative models. Contemporary linguistic models are now capable of synthesizing text with remarkable fluency[33, 35, 41], while novel text-to-image models have demonstrated the ability to fabricate images of striking realism[4, 28]. These advancements herald a new era for the pragmatic deployment of artificial intelligence across various sectors, including artistry, design, and the broader spectrum of content generation.

The creation of high-quality visual story of indeterminate duration continues to be a formidable challenge within the field of artificial intelligence. Unlike the generation of a single image, the synthesis of the multi-frame visual story requires the maintenance of subject continuity across a sequence of frames, which poses a considerable obstacle. Within this context, the task of producing visually stunning images with storytelling attributes is particularly arduous. Nevertheless, recent innovations in artificial intelligence have given rise to models that demonstrate adeptness in crafting videos with coherent and engaging narratives[27, 31].

Although a large number of video datasets [2, 37, 40] exist, capable of providing sufficient training data for story visualization models, these datasets lack subject focus. Moreover, the ability of image sequence to narrate a story is relatively limited, as evidenced by the absence of subtitles and thematic masks. Additionally, many story visualization datasets are confined to a closed domain[14, 16, 17, 21], which is highly unfavorable for the story generation model to generate more naturalistic stories. Consequently, to enhance the story generation capabilities of storytelling image generation models and reduce the associated costs of model training, higher-quality, subject-focused open-domain story visualization datasets, and low-cost dataset production methods are urgently required.

As a result, our research outlines a comprehensive pipeline designed to supply training data for visual story generation models, emphasizing the maintenance of narrative continuity around pivotal subjects and representing the dataset **OpenStory**, cleaned through this pipeline. To enhance the quality and cost-effectiveness of the dataset, our pipeline employs an efficient approach 1 to extract salient keyframes from the video content. We leverage an aesthetic evaluation model to filter out high-quality keyframes, and subsequently utilize BLIP2[12], a powerful vision-language model, to generate captions for these keyframes.

Moreover, to enhance continuity between the keyframes and captions, we harness the power of a Large Language Model (LLM) to optimize the captions, ensuring that they smoothly depict a coherent narrative. This refinement process is conducive to training story visualization models that specialize in storytelling. Furthermore, to deliver higher-quality data, our sub-pipeline incorporates the capability to identify valid subjects within the images and generate corresponding subject masks using the Segment Anything Model (SAM)[11].

Upon the completion of this comprehensive pipeline, we have successfully transformed video content into subject-focused image sequences, culminating in the creation of the **OpenStory** Dataset. Unlike other datasets, **OpenStory** is uniquely positioned to provide nourishment for subject-focused, storytelling-centric story visualization generation models, as it encapsulates narratively coherent visual sequences tailored to the specific subjects of interest. Next, we will show more details of our pipeline and dataset in Section 3.

# 2. Related Works and Motivations

## 2.1. Text-to-image generation

A multitude of studies have showcased remarkable advancements in single-image generation. Recently, diffusion-based text-to-image models[4, 28, 29] have demonstrated significant progress in enhancing image quality and diversity by leveraging diffusion models. However, these text-to-image approaches predominantly focus on aligning individually generated images with textual descriptions, neglecting the crucial aspects of character and scene consistency across multiple frames in the story visualization task. Moreover, they struggle to effectively resolve ambiguous references within narrative descriptions. To address this issue, we introduced a Language-Image Model (LLM) to optimize the captions of image sequences, thereby ensuring overall coherence across the entire sequence. Furthermore, the field of customized diffusion-based models has witnessed considerable activity of late. Subject-focused controllable T2I generation models like IP-Adapter[39], PhotoMaker[15], Break-A-Scene[1], and others have garnered widespread attention. And compositional generation works like TF-COM[7], R3CD[18], AMC[36]. We aim to extend this capability to story visualization tasks, which can also be effectively applied to other tasks such as subject-focused story visualization. Therefore, our pipeline can transform continuous frame sequences into frames with subjects containing masks and bounding boxes.

## 2.2. Multimodal Datasets

The effectiveness of crossmodal learning hinges on the harmonious integration of visual and textual data. To develop robust vision-language representations, datasets must not only be voluminous but also exhibit strong visual-textual alignment. In academia, the use of images with associated alternative text[6, 9, 10] and videos with ASR transcriptions[25, 32, 38]to achieve scalable learning. The advent of LAION-5B [30] has provided access to an unparalleled number of image-text pairs, significantly advancing the field of image-language pretraining. In the sphere of video-centric multimodal learning, the HowTo100M dataset[25] has been pivotal, leveraging ASR subtitles from instructional videos to promote the learning of joint representations. Additionally, the YT-Temporal[40] has been dedicated to the fusion of audio-visual-language elements and the crossmodal learning of high-resolution videos. Bain[3] has highlighted the criticality of accurate video-text alignment, surpassing the need for large data quantities, which led to the inception of the WebVid dataset[2], now a cornerstone in modern video-language pretraining strategies[13]. Therefore, the need for an effective pipeline to make high-quality supply story visualization models is urgent.

## 2.3. Story Visualization and Story Dataset

### 2.3.1 Story Visualization model

Pioneering the story generation task, StoryGAN[14] proposed a sequential conditional GAN framework with dual frame and story level discriminators to enhance image quality and narrative coherence. Then two GAN-based improvements, DuCo-StoryGAN[23] and VLCStoryGAN[22] utilizing video captioning for semantic alignment between text and frames, further improved the model's tory-generation capabilities. After that, StoryDALL-E[24] further refined this approach by retrofitting the cross-attention layers of a pre-trained text-to-image model to enhance generalizability to unseen visual attributes in generated stories. However, these methods often overlook ambiguous references in text descriptions. StoryLDM[27] addressed this gap by introducing reference resolution in story visualization tasks, employing an autoregressive fusion-based framework with a memory-attention module. Besides, StoryGPT-V[31] introduces a two-stage approach combining latent diffusion models and large language models to generate consistent and high-quality characters for visual story visualization by resolving anaphora and leveraging reasoning abilities. To further unlock the potential of story generation models, it is essential to have a high-quality and continuous frames sequence-captions pairs dataset. Having a pipeline that can easily derive such data from open-domain video data would be highly beneficial.

### 2.3.2 Story Visualization Dataset

While exist numerous datasets suitable for training subject-focused storytelling video generation models, they encounter several limitations. The earliest VIST[34] dataset, for instance, lacked subject extraction within frames, hindering effective subject-focused story visualization tasks. Additionally, datasets like PororoSV[14], FlintstonesSV[21], and StorySalon[17] offer better continuity in frame sequence and caption coherence. However, they are confined to close-domain scenarios such as cartoons, thus constraining the model's capability in open-domain scene generation. Furthermore, artificial annotations in captions of datasets like PororoSV, FlintstonesSV, and VIST restrict dataset size and escalate production costs. Conversely, datasets like DideMoSV[24] rely on video subtitles for caption generation, potentially compromising caption accuracy in describing frame scenes. To address these limitations, we devised a pipeline shown in figure Figs. 1 and 2 capable of extracting subject-focused keyframes and sequences from open-domain videos. Utilizing captions with exceptional continuity, we introduce the **OpenStory** dataset, a large-scale dataset derived from this pipeline. Our dataset comprises caption-frame pairs with long-term sequences. We anticipate that this dataset will significantly enhance the model's proficiency in subject-focused story visualization tasks. Moreover, our dataset, featuring caption-frame pairs with long-term sequences, establishes a robust benchmark for the challenging task of multi-modal long-context understanding.

## 3. OpenStory

Seeking to equip subject-focused story visualization models with abundant training data, we introduce OpenStory, a dataset centered on the continuous frames of subjects, aimed at facilitating the synthesis of coherent and contextually relevant visual narratives. In contrast to earlier methodologies that conceptualize the task of story visualization as generating a solitary keyframe in response to each textual cue, our innovative approach redefines this paradigm, our pipeline excels at isolating key frames from any given video and adeptly concentrates on the subject as they undertake various actions. By leveraging a meticulous frame extraction process that emphasizes the most significant segments, our methodology ensures the substantial facets of the video are captured, enabling a focused analysis of the subject's activities.

The intricacies of video generation for cutting-edge models are multifaceted and demanding. The pivotal challenges encompass the creation of videos that (i) unfold with a seamless and logical narrative, (ii) manifest with lifelike visual fidelity, and (iii) adhere to the directives set forth by user preferences Our work advances video generation by creating a pipeline that (i) efficiently extracts key frames from large video datasets, (ii) enriches narratives by focusing on varied actions, and (iii) combines computer vision with natural language processing for coherent frame selection and description. This approach not only refines training data but also improves the storytelling quality of generated videos, ensuring they are both visually appealing and contextually rich.

### 3.1. Datasets

We reference and process mainstream video datasets such as InternVid[37], StorySalon[25], and Howto100M[25], among others. We download a substantial number of videos from YouTube just like them and incorporate data from the mentioned datasets that align with the requirements of our subject-focused story visualization task into our dataset.

For instance, the InternVid dataset encompasses 18 million videos annotated by a video understanding model integrating text-image alignment models and large language models. Despite containing a subset of 10 million videos curated by an aesthetic evaluation model, many fail to meet quality standards for subject-focused story visualization, as they lack clear narratives and distinct subjects. Short video slices also present challenges in extracting high-quality frames. Therefore, we exclusively utilize full videos from InternVid, filtering them based on paired subtitles and aesthetic scores to ensure thematic relevance and quality. Additionally, we incorporate appropriate fragments from close-domain datasets such as StorySalon and PororoSV to enhance dataset generalization and indirectly improve the story visualization model's generalization capability.

This calibrated inclusion strategy allows us to maintain the dataset's narrative strengths while simultaneously broadening the learning horizon of our models, enabling them to generate a more varied and representative array of video content.

### 3.2. Pipeline

The progression of sophisticated text-to-video modeling faces a significant obstacle due to the scarcity of substantial, high-quality datasets. The manual annotation of video content to construct narrative-driven formats is widely acknowledged as a labor-intensive and costly endeavor. This challenge becomes particularly daunting when aiming to capture the temporal dynamics and story elements crucial for training such models. Consequently, there is a pressing need for automated and scalable data curation methods to address these challenges and enhance the training process

Table 1. OpenStory statistics and comparison with other story-visualization datasets and in OpenStory, more than 80% of the frames come from 720P and 1080P videos. ”-” means it is not easy to collect or this dataset lacks this indicator.

| Dataset | Domain | Caption | Frames | Avg. Length | Resolution | Masks per frames | seriality |
|---|---|---|---|---|---|---|---|
| **PororoSV[14]** | close | Manual | 73K | 5 | - | - | ✓ |
| **FlintstonesSV[21]** | close | Manual | 123K | 5 | - | - | ✓ |
| **DideMoSV[24]** | close | Manual | 53K | 3 | - | - | ✓ |
| **VIST[34]** | open | Manual | 145K | 5 | - | - | ✓ |
| **StorySalon[17]** | close | ASR | 159K | 14 | - | 1 | ✓ |
| **SDD[20]** | open | Generated | 76M | 1 | - | 3 | ✗ |
| **OpenStory (ours)** | open | Generated | 107M | 28 | 720P+1080P | 2.5 | ✓ |

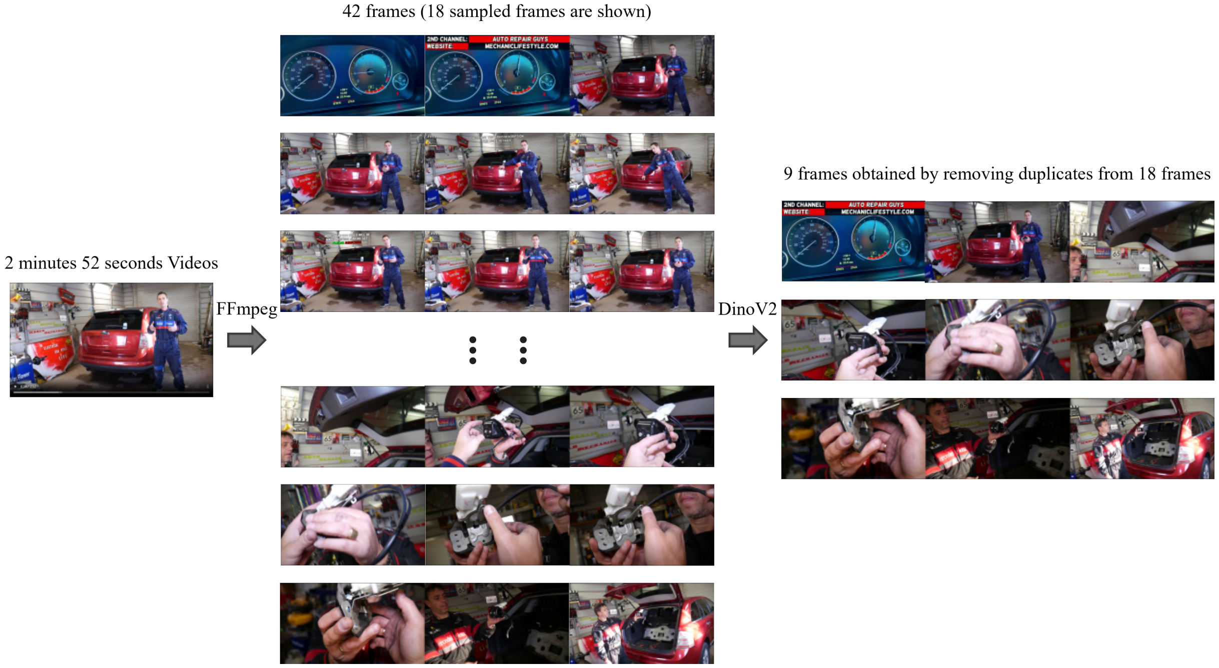
42 frames (18 sampled frames are shown)
9 frames obtained by removing duplicates from 18 frames
2 minutes 52 seconds Videos
FFmpeg
DinoV2

Figure 1. We employ FFmpeg[8] for the extraction of I-frames from the video, followed by the utilization of DinoV2[26] for frame encoding. Subsequent frames are compared, and if the cosine similarity surpasses a predefined threshold, the antecedent frame is eliminated to ensure optimal data compression.

of these models.

Due to the above reasons, in this subsection, we delve into the details of our pipeline, a multi-faceted process that seamlessly transitions from video content to keyframes, refines these keyframes through deduplication, and then employs these frames to generate descriptive captions. Furthermore, we utilize a highly sophisticated Large Language Model (LLM) to refine these captions to near perfection. Finally, the pipeline harnesses the synergy of keyframes and their polished captions to create precise subject masks. The pipeline of our work is shown in Figure Figs. 1 and 2

**Single-image captioning and subject-masking** The workflow for processing a single image begins with blip2[12] which leverages a Language Model (LLM) capable of executing zero-shot image-to-text conversion guided by natural language prompts automatically adding captions to the image. Subsequently, NLTK[5] is employed for subject segmentation based on the generated captions. Once the subject is identified, Grounding-dino[19] locks its bounding box and feeds it into SAM[11] to produce a pixel-level subject mask. This entire process is fully automated and operates independently of large-parameter models.

**Extracting Frames** The initial phase of our frame extraction pipeline focuses on identifying and isolating I-frames (intra-coded frames) from the video content. I-frames serve as the foundation of video compression and act as self-contained reference points that can be decoded independently, without relying on information from preceding or

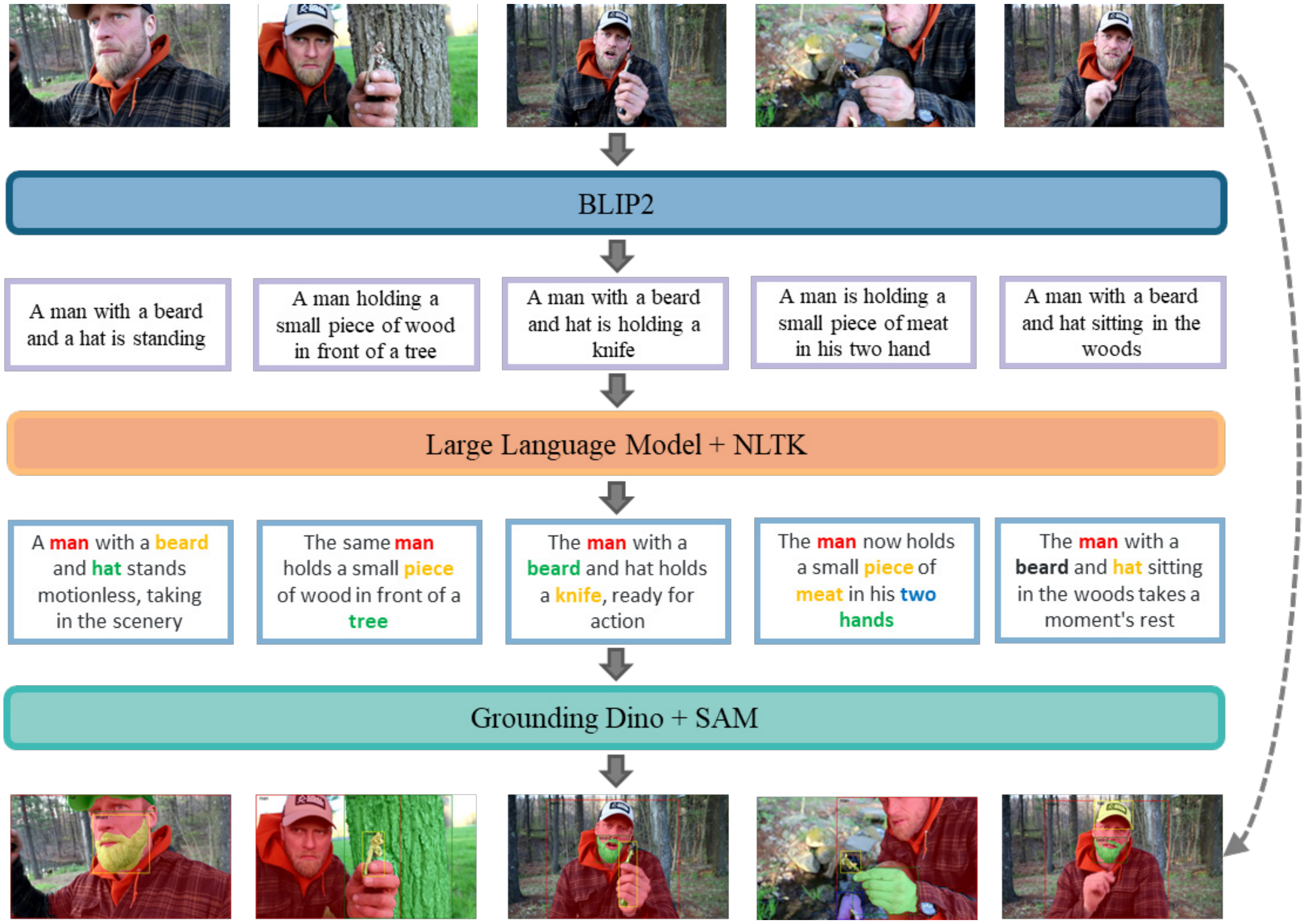
BLIP2
A man with a beard and a hat is standing
A man holding a small piece of wood in front of a tree
A man with a beard and hat is holding a knife
A man is holding a small piece of meat in his two hand
A man with a beard and hat sitting in the woods
Large Language Model + NLTK
A man with a beard and hat stands motionless, taking in the scenery
The same man holds a small piece of wood in front of a tree
The man with a beard and hat holds a knife, ready for action
The man now holds a small piece of meat in his two hands
The man with a beard and hat sitting in the woods takes a moment's rest
Grounding Dino + SAM

Figure 2. This figure shows the pipeline that upon acquiring the de-duplicated keyframes, we employ the BLIP2[12] model to generate initial captions for each frame. Subsequently, LLM[41] is utilized to refine these captions, ensuring narrative continuity across the sequence of frames. The captions are then extracted using the NLTK[5]. To determine the subject's position within the image, we apply the Grounding Dino model, and for the segmentation of the subject, the SAM[11] model is used to produce the corresponding masks.

subsequent frames. This inherent characteristic makes I-frames the ideal candidates for keyframe extraction. Upon successfully extracting these I-frames, we leverage the capabilities of DINOv2[26] to extract their Vision Transformer features. This advanced technique enables us to identify frames exhibiting high visual similarity. Specifically, if the cosine similarity between the feature vectors of consecutive frames surpasses a predetermined threshold, indicating redundancy, we strategically eliminate the preceding frame. This approach ensures the generation of a diverse and concise set of keyframes, optimizing the efficiency and effectiveness of subsequent processing steps. This optimized selection of keyframes, delineating the reduction of redundancy and enhancement of diversity, is visually summarized in Figure 1, which illustrates the overarching frame extraction pipeline.

**Frame-Caption Alignment** With the extracted keyframes at our disposal, our focus shifts to the crucial task of aligning these frames with their corresponding textual descriptions. Similar to the approach used for single-image processing, we leverage the BLIP-2 model, which acts as a bridge between the visual and linguistic domains, facilitating the seamless integration of visual and textual information. Once initial captions are generated by BLIP-2, we design a prompt, as illustrated in Figure 3, and employ ChatGLM3-Turbo[41], an advanced large language model, to implicitly deduce the subject of references. ChatGLM3-Turbo is specifically tasked with maintaining consistency in subject matter across captions, ensuring a coherent and seamless narrative flow, even during transitions between different scenes. The LLM plays a pivotal role in ensuring that the language used in the captions is not only descriptive but also maintains consistent subject con-

Refine captions for improved narrative flow and coherence across scenes. Enrich imagery with descriptive details without altering core content. Maintain consistent subjects throughout.

Guidelines:

1. Smooth phrasing and sentence structure for better storytelling
2. Develop a unified narrative arc spanning multiple scenes
3. One-to-one correspondence in input/output formatting
4. Keep optimized captions concise while enhancing the narrative
5. In captions, nouns with similar meanings that represent the subject should use the same one.
6. The nouns used should be as simple as possible and do not use complex nouns.
7. Never include the input caption in your reply!
8. You response must be and only have a list containing the length of the input list of captions and subject consistency which are most important, you must obey!

example input: ["People gather in the park, enjoying a sunny day.", "People throw frisbees, laughter echoing.", "People set up a picnic, blankets spread on the grass."]
example output: ["People congregate in the verdant park, reveling under the warm sun.", "The same people engage in cheerful frisbee games, their laughter mingling with the breeze.", "These people then prepare a picnic, arranging blankets neatly upon the lush grass."]

My Input Captions: ...

Figure 3. This figure presents a prompt designed to enhance narrative flow and coherence across scenes, which contains refined captioning guidelines aimed at enriching imagery with descriptive details while preserving the core content. Additionally, the prompt emphasizes maintaining consistent subjects throughout the storytelling process.

tinuity throughout the video. This consistency is essential for subject-focused story visualization tasks, as it facilitates the generation of cohesive and engaging narratives.

**Frames to Masks for Subject-Driven Task** Story-telling continuous keyframe processing is very similar to the subject masking process for a single image. Initially, we utilize NLTK to conduct subject segmentation based on the caption, followed by employing Grounding Dino to pinpoint the subject's position within the frame. Subsequently, SAM is utilized to execute precise pixel-level masking. However, a key distinction from the single-image processing flow lies in the emphasis on maintaining subject consistency across frames for our subject-focused story visualization task. To achieve this, we leverage a LLM to refine captions, ensuring that they accurately reflect ongoing narratives before proceeding with subject segmentation and maintain subject consistency between captions. This preliminary step optimizes the coherence and continuity of subjects throughout the video, laying the groundwork for precise image-masking within each frame.

**Discussion** Our automatic dataset-generation pipeline has demonstrated considerable efficacy in producing high-quality datasets tailored for subject-focused story visualization tasks. By seamlessly integrating advanced language and vision models, we have successfully automated the extraction of keyframes from video content and aligned them with corresponding textual descriptions. This pipeline ensures the creation of datasets that not only exhibit narrative coherence but also maintain subject continuity across frames, essential for effective story visualization. The datasets produced through this pipeline have the potential to serve as a benchmark for multi-modal long-context understanding tasks, facilitating the training of models capable of comprehensively interpreting and generating narratives from extensive visual and textual inputs. Moving forward, our pipeline offers a scalable and efficient approach to dataset creation, enabling researchers to access rich and diverse datasets that can significantly advance the field of subject-focused story visualization and multi-modal understanding.

### 3.3. Dataset Building Distribution

In comparison to other datasets outlined in Table 1, our dataset possesses several distinct advantages. Firstly, it embraces an open-domain approach, ensuring its applicability across diverse scenarios and subjects. Secondly, our dataset benefits from fully automated captioning processes, eliminating the need for manual annotation and significantly reducing the associated labor costs. Moreover, it boasts a large-scale corpus, providing ample training data for robust model development. Additionally, our dataset features long frame sequences, enabling the exploration of temporal dynamics and complex narratives. Furthermore, the dataset

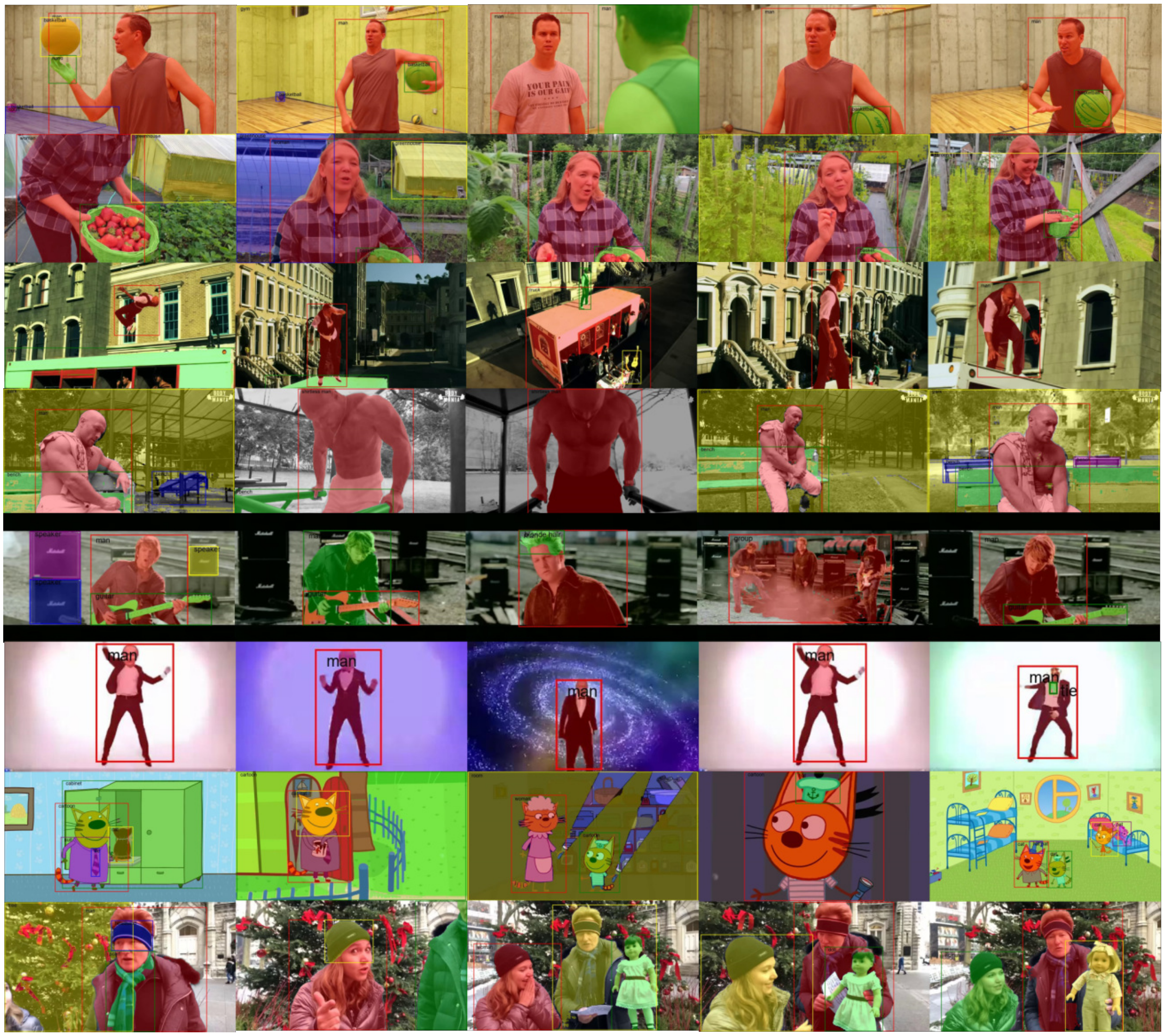

Figure 4. This figure shows the video frame sequence generated by our pipeline, with the subject's mask and the subject's bounding box.

offers high-resolution imagery, ensuring detailed visual representations. Importantly, it strikes a balance by including just the right number of subjects within each frame, avoiding clutter, and maintaining focus. Lastly, our dataset excels in ensuring continuity between frames, facilitating seamless transitions, and enhancing the overall narrative coherence. These collective advantages position our dataset as a valuable resource for advancing research in subject-focused story visualization and multi-modal understanding.

Statistics about our proposed dataset are illustrated in Figure 5, which presents a comprehensive analysis of the dataset properties, including the distribution of resolution, aesthetic score, caption length, and subject number per keyframe. The dataset emerges as a paragon of high-resolution, subject-focused video keyframes, meticulously annotated with captions that encapsulate the essence of each frame.

The resolution distribution reveals that an overwhelming 83.4% of keyframes are rendered in the pristine clarity of 1080p, with a substantial 11.8% at 720p, heralding a dataset resolutely committed to delivering high-definition visual acuity. Conversely, the presence of resolutions beneath 360p is virtually non-existent, a mere 1.1%, underscoring the dataset's fidelity to superior image quality.

Moreover, the aesthetic caliber of the dataset is rigorously quantified, with a notable 43.5% of keyframes achieving an aesthetic score exceeding 6, indicative of a collection curated with an astute appreciation for visual excellence.

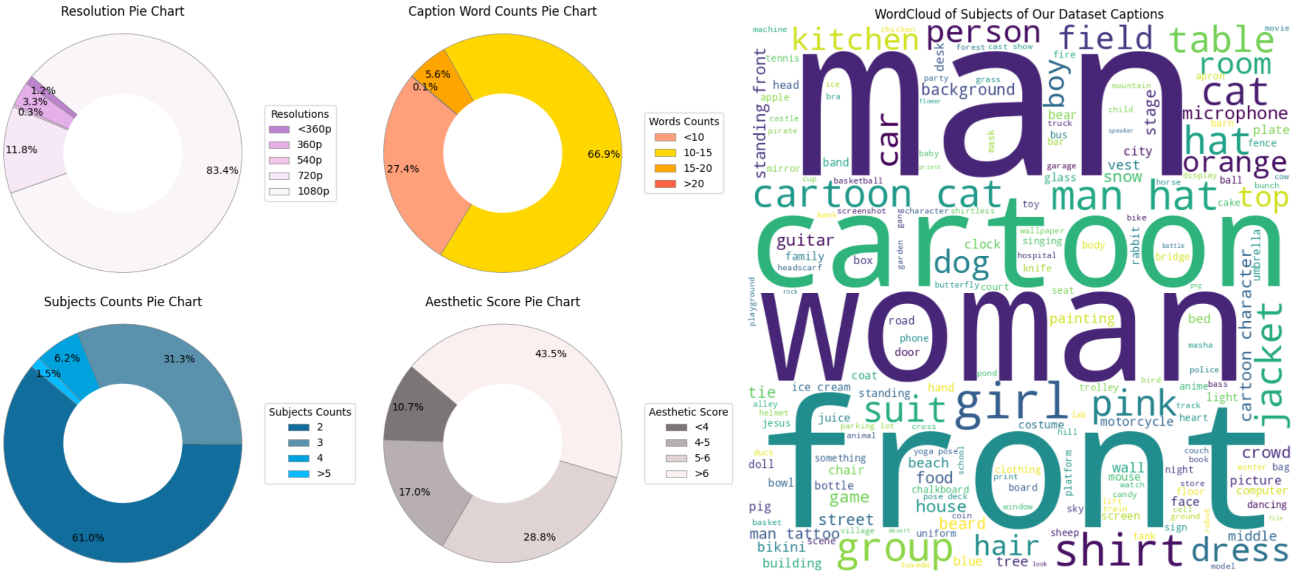

Figure 5. Statistics for our dataset. The left part is the composition ratio of each evaluation indicator in our dataset; The right part is the word cloud distribution map of the elements in our dataset.

The gradation of visual allure is further delineated, with 28.8% scoring between 5-6 and 17.0% between 4-5, while a modest 10.6% falls below the threshold of 4, showcasing the dataset's encompassing range of visual aesthetics.

In addition to the visual quality, the caption length distribution demonstrates the dataset's dedication to conciseness without compromising descriptive depth. A significant 66.9% of captions are succinctly confined to 10-15 words, while captions of fewer than 10 words account for 27.3%, offering crisp synopses. Captions exceeding 15 words are a rarity, with 5.6% within 15-20 words and a scant 0.07% surpassing 20 words, mirroring the dataset's emphasis on relevance and succinctness.

Furthermore, the subject number distribution accentuates the dataset's aptitude for scenarios that demand a concentrated focus on individual or paired subjects, eschewing the complexity of densely populated scenes. The dataset's composition is rich in dual-subject scenarios, with 59.6% of keyframes featuring precisely two subjects. Tri-subject frames are also prevalent, constituting 30.6%, while instances of four and more than five subjects are comparatively sparse, at 6.0% and 1.4% respectively.

Our dataset stands out for its open-domain approach, fully automated captioning, large-scale corpus, long frame sequences, high-resolution imagery, balanced subject composition, and seamless narrative coherence. These features position our dataset as a valuable resource for advancing research in subject-focused story visualization and multi-modal understanding.

# 4. Conclusion

In conclusion, we presents OpenStory, a large-scale, high-quality dataset that enhances the training of image generation based subject-focused story visualization models. This dataset emerges from a robust pipeline that harnesses the synergy of advanced computer vision and natural language processing techniques. It adeptly extracts keyframes from a large amount of open-domain videos, annotates them with descriptive captions, and refines these captions for narrative coherence through sophisticated large language models, culminating in the creation of precise subject masks. OpenStory offers the following advantages: (1) its open-domain nature ensures versatility across a myriad of subjects and scenarios; (2) its automated caption generation system obviates the need for labor-intensive manual annotation; its extensive collection boasts millions of frame-caption pairs; (3) it supports the examination of extended temporal narratives through long sequences; (4) it provides high-resolution imagery for impeccable clarity; (5) it maintains a focus on subjects to minimize extraneous visual elements; (6) and it ensures frame-to-frame continuity for seamless storytelling. As a benchmark for subject-focused story visualization and long-context multi-modal understanding, OpenStory is set to revolutionize research in these domains. The pipeline itself is a testament to efficiency and scalability, offering a streamlined approach to curating rich multi-modal datasets from the vast expanse of open video sources. The contributions of this work, particularly in dataset creation, are poised to catalyze a new wave of innovation in narrative story visualization and expand the horizons of multi-modal content creation.

## References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*. ACM, 2023. 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 1, 2

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2022. 2

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions, 2023. 1, 2

[5] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009. 4, 5

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021. 2

[7] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 2

[8] FFmpeg Developers. Ffmpeg. https://github.com/FFmpeg/FFmpeg, 2024. 4

[9] Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models, 2023. 2

[10] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning, 2022. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 4, 5

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 2023. arXiv:2301.12597 [cs]. 2, 4, 5

[13] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models, 2024. 2

[14] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization, 2019. 1, 2, 3, 4

[15] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023. 2

[16] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning, 2023. 1

[17] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent Grimm – Open-ended Visual Storytelling via Latent Diffusion Models, 2024. arXiv:2306.00973 [cs]. 1, 3, 4

[18] Jinxiu Liu and Qi Liu. R3cd: Scene graph to image generation with relation-aware compositional contrastive control diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3657–3665, 2024. 2

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[20] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-Diffusion:Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning, 2023. arXiv:2307.11410 [cs]. 4

[21] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization, 2021. 1, 3, 4

[22] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization, 2021. 2

[23] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency, 2021. 2

[24] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation, 2022. 2, 3, 4

[25] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, 2019. 2, 3

[26] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4, 5

[27] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation, 2023. 1, 2

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 2

[31] Xiaoqian Shen and Mohamed Elhoseiny. Storygpt-v: Large language models as consistent story visualizers, 2023. 1, 2

[32] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2021. 2

[33] Gemini Team. Gemini: A family of highly capable multimodal models, 2023. 1

[34] Ting-Hao, Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual Storytelling, 2016. arXiv:1604.03968 [cs]. 3, 4

[35] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. 1

[36] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*, 2023. 2

[37] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation, 2024. arXiv:2307.06942 [cs]. 1, 3

[38] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions, 2022. 2

[39] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 2

[40] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *CVPR*, 2022. 1, 2

[41] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 1, 5