FlintstonesSV++ : Improving Story Narration using Visual Scene Graph

Janak Kapuriya[1,*], Paul Buitelaar[1]

[1]*Insight Research Ireland Centre for Data Analytics, Data Science Institute, Univeristy of Galway, Ireland*

Abstract

Recent advancements in text-to-image, text-to-video, and large language models have significantly enhanced the performance of various downstream tasks. In the field of Story Visualization, models have been developed to generate coherent image sequences from storylines composed of multiple scenes. These innovations have largely relied on benchmark datasets such as FlintstonesSV and PororoSV, which provide essential resources for tasks like Story Visualization and Story Continuation. However, our analysis identifies several limitations in the FlintstonesSV dataset that restrict the performance of models trained on it. To address these limitations, we introduce FlintstonesSV++, an enhanced version of the FlintstonesSV dataset. FlintstonesSV++ leverages visual Scene Graphs and Large Language Models to enrich storylines with factual details, further validated by human reviewers. By fine-tuning text-to-story generation models on FlintstonesSV++, we demonstrate substantial improvements, achieving a 5.2% average increase in alignment scores and a 5.72% boost in image generation quality compared to models trained on the original dataset. Moreover, a qualitative comparative analysis highlights the superior performance of FlintstonesSV++ compared to the original dataset. The FlintstonesSV++ dataset marks a significant advancement in enabling tasks such as Story Visualization and Story Continuation. To support further research in story-based visual content generation, we made the code and dataset publicly available.

Code and Dataset - https://github.com/janak11111/FlintstonesSV++

Keywords

Story Narrative Generation, Visual Scene Graphs, Dataset Improvement, Storyline Visualization, Narrative Resources, Large Language Models, Large Multimodal Models

1. Introduction

Recent advancements in text-to-image generation have been driven by high-quality models such as DALL-E 3 [1], SDXL [2], Imagen-3 [3], and Stable Diffusion-3 [4]. These models excel in generating photorealistic images, which has advanced the development of text-to-video generation models like Sora [5], SVD [6], and Veo2 [7]. Real-world applications of these models include personalized product advertisements, story visualization, educational content creation and Social media video creation. Among these applications, story visualization transforms textual narratives into coherent visual representations, bridging the gap between language and visual understanding. A related task, story continuation, generates coherent image sequences based on a given storyline and an initial image. Unlike traditional text-to-image or text-to-video generation, which typically focus on isolated descriptions, story visualization and continuation require generated images to align with textual sentences while maintaining consistency in characters and scenes. These tasks are further complicated by challenges such as ensuring frame-to-frame coherence, capturing context, and addressing factual gaps in scene descriptions.

Recent progress in the story visualization domain highlights promising approaches. Grimm [8] employs an auto-regressive method to establish temporal dependencies between image-sentence pairs. Story-LDM [9] uses a visual memory module to maintain consistency across generations, while ARLDM

In R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10-April-2025

[*]Corresponding author.

janakkumar.kapuriya@insight-centre.org (J. Kapuriya); paul.buitelaar@universityofgalway.ie (P. Buitelaar)

https://janak11111.github.io/ (J. Kapuriya);

https://www.universityofgalway.ie/our-research/people/computer-science/paulbuitelaar/ (P. Buitelaar)

0009-0007-9562-7672 (J. Kapuriya); 0000-0001-7238-9842 (P. Buitelaar)

[10] applies text-to-image diffusion models for coherent image generation. Temporal-Story [11] integrates a flow adapter and spatio-temporal attention to capture character movements across scenes. These models have been evaluated on benchmark datasets like FlintstonesSV [12] and PororoSV [13], designed for animated story visualization tasks. Despite their utility, the FlintstonesSV dataset has limitations that hinder model performance in generating coherent and consistent scenes.

Each sample in FlintstonesSV includes an image and a corresponding scene description, typically covering basic details like the character's name, activity, and setting. However, key elements essential for a comprehensive understanding of the scene are often missing. Specifically, it lacks essential details, such as character attributes, detailed background descriptions, precise character positioning, and high-level objects along with their relationships to other objects and characters. Addressing these limitations is vital for advancing story visualization and continuation tasks.

To address the limitations of the FlintstonesSV dataset, we propose a Visual Scene Graph (VSG)-based approach to enhance the factual accuracy of its scene descriptions. A VSG [14] extracts key information from images, such as objects, attributes, and relationships, which serves as a foundation for generating enriched scene descriptions using large language models. This process results in the enhanced FlintstonesSV++ dataset, which significantly improves the performance of text-to-image generation models in terms of alignment and generative quality. We validate our improvements through expert evaluations of the visual scene graphs and provide qualitative comparisons between the original FlintstonesSV and FlintstonesSV++ descriptions. Additionally, results from text-to-story generation models fine-tuned to generate scene images from scene narratives further highlight the effectiveness of our approach.

Our key contributions in this paper are:

1. **Visual Scene Graph Integration:** We introduce a novel approach that leverages Visual Scene Graphs to address the limitations of the FlintstonesSV dataset, enriching it with detailed factual information, which is subsequently validated by human reviewers.
2. **FlintstonesSV++ Dataset Creation:** We develop an improved version of the FlintstonesSV dataset, named FlintstonesSV++. This enhanced dataset combines Visual Scene Graphs with large language model outputs to create more comprehensive and accurate scene descriptions.
3. **Performance Improvements :** Our experiments with FlintstonesSV++ demonstrate significant enhancements in model performance. We observe an average 5.2% increase in scene description alignment CLIP score and average 5.72% improvement in text-to-story generation results across various pretrained diffusion models.
4. **Qualitative Comparative Analysis :** We provide an in-depth qualitative comparison between FlintstonesSV++ and the original FlintstonesSV dataset. This analysis highlights the superior quality of our improved dataset, showcasing enhanced scene descriptions and more accurate prediction results.

2. Related Work

2.1. Story Visualization

The Story Visualization task focuses on generating sequences of visually coherent scenes from multi-scene storylines. Earlier approaches predominantly relied on Generative Adversarial Networks (GANs) to generate scene sequences [15, 16, 17, 18, 19, 20]. More recently, diffusion models have been introduced for this task, showing promising results [9, 21, 10, 22, 11]. These methods often utilize common benchmark datasets such as FlintstonesSV [12] and PororoSV [13]. However, we identified inherent issues with the FlintstonesSV dataset. It lacks comprehensive background information, precise character attributes, accurate positioning of characters, and detailed object descriptions, including their relationships with other objects and characters. These limitations hinder the performance of downstream tasks, such as story visualization, particularly when fine-tuning story visualization models.

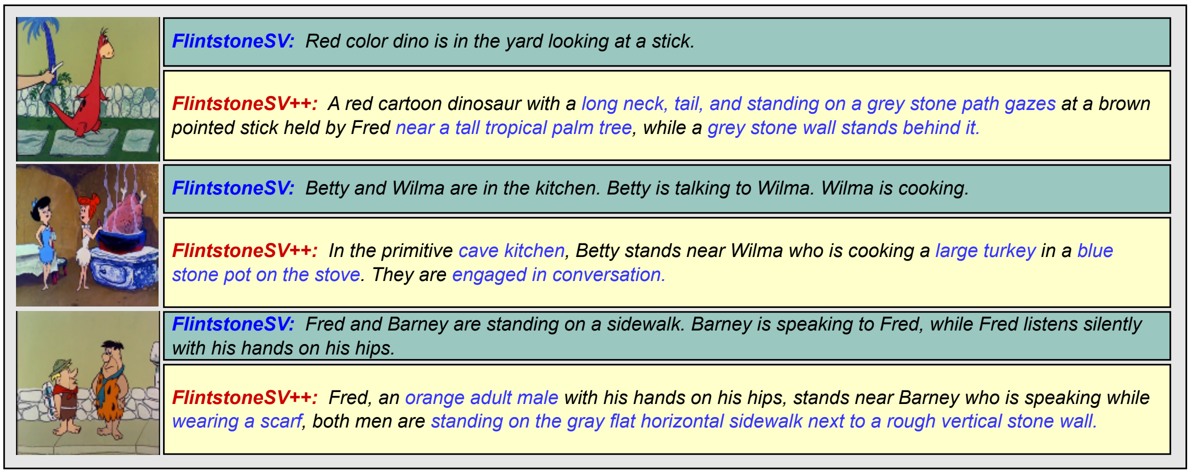
2.2. Visual Scene Graphs

The Visual Genome framework [14] extracts structured factual information from images by identifying objects, attributes, and relationships within distinct image regions. These region-specific graphs are combined into a unified scene graph, where objects are interconnected through specific relationships and annotated with relevant attributes, effectively integrating factual information from images. Scene graphs have shown substantial utility in various downstream tasks, including Visual Question Answering [23, 24, 25], Visual Scene Reasoning [26, 27, 28], and Image Captioning [29, 30], leading to notable improvements in task accuracy. In our work, we adopt a Visual Scene Graph (VSG) based approach to address inherent issues in the original FlintstonesSV dataset. By utilizing VSGs, we aim to improve the dataset's quality by addressing gaps such as incomplete background information, imprecise character attributes, and inadequate object relationships.

3. Dataset

3.1. FlintstonesSV Dataset and its Limitations

The FlintstonesSV dataset is curated from the animated sitcom *The Flintstones*, featuring scenes centered around seven main characters, each contributing to diverse interactions and scenarios. These characters play a pivotal role in capturing the dynamics and relationships within the series. The dataset is composed of 24,512 samples, divided into training, validation, and test sets comprising 20,132, 2,071, and 2,309 samples, respectively. Each sample consists of an image paired with a concise scene description. The FlintstonesSV dataset has been widely used for tasks such as story visualization and story continuation. Its alignment of images with corresponding scene descriptions makes it a valuable resource for developing and evaluating models that integrate visual content with narrative generation.



FlintstoneSV: *Red color dino is in the yard looking at a stick.*

FlintstoneSV++: *A red cartoon dinosaur with a long neck, tail, and standing on a grey stone path gazes at a brown pointed stick held by Fred near a tall tropical palm tree, while a grey stone wall stands behind it.*

FlintstoneSV: *Betty and Wilma are in the kitchen. Betty is talking to Wilma. Wilma is cooking.*

FlintstoneSV++: *In the primitive cave kitchen, Betty stands near Wilma who is cooking a large turkey in a blue stone pot on the stove. They are engaged in conversation.*

FlintstoneSV: *Fred and Barney are standing on a sidewalk. Barney is speaking to Fred, while Fred listens silently with his hands on his hips.*

FlintstoneSV++: *Fred, an orange adult male with his hands on his hips, stands near Barney who is speaking while wearing a scarf, both men are standing on the gray flat horizontal sidewalk next to a rough vertical stone wall.*

Figure 1: Comparison of scene narratives from the FlintstonesSV and FlintstonesSV++ datasets. The text highlighted in blue in FlintstonesSV++ represents additional factual information extracted through our visual scene graph-based approach, which is missing in the original FlintstonesSV dataset

Despite its utility, the FlintstonesSV dataset has notable limitations. As illustrated in Figure 1, Example 1 highlights that FlintstonesSV captions fail to provide information about Dino and background elements, such as the wall and palm trees. In Example 2, the original captions omit crucial details, including the type of food being cooked, the utensil used, and its color. Similarly, in Example 3, the captions neglect to describe character apparel, such as a scarf, and background elements like the wall, along with their spatial relationships with the character. These gaps limit the dataset's ability to capture the complete essence of a story scene. Consequently, models trained on it often struggle with generating or continuing stories that are contextually rich and detailed. The lack of critical visual information in scene descriptions further hampers performance in tasks such as story generation and continuation.

3.2. Narrative Improvement Using Visual Scene Graphs

Inspired by the Visual Genome framework [14], we adopt a similar methodology to enhance text-to-story generation for the FlintstonesSV dataset. By utilizing a visual genome-based approach, detailed information about story scenes can be extracted from the FlintstonesSV dataset, capturing key factual details enriches the scene representation, addressing the limitations of the FlintstonesSV dataset. This added detail is pivotal for story visualization and text-to-story generation, as it provides a deeper understanding of the narrative context. Consequently, models can generate more coherent, contextually accurate, and visually grounded stories, significantly improving the quality of their outputs.

Figure 1 showcases the effectiveness of the FlintstonesSV++ dataset by comparing scene images and their corresponding captions from the original FlintstonesSV dataset with the enhanced captions of FlintstonesSV++. Missing details in the FlintstonesSV dataset are accurately captured in FlintstonesSV++ and highlighted in blue. This qualitative analysis highlights the significant improvements made in FlintstonesSV++, addressing critical gaps in the original dataset. These enhancements are achieved through factual information extracted using the visual scene graph, resulting in a more comprehensive and factually enriched dataset.

3.3. Visual Scene Graph Human Evaluation

To validate the accuracy of the Visual Scene Graphs (VSGs) extracted from story scenes, we conducted a human evaluation with 7 annotators, all of whom are researchers in NLP, and of which 3 are PhD students, 3 are research staff and 1 is academic staff. Each evaluator reviewed 10 randomly selected VSG samples. The evaluation focused on three primary components of the generated VSGs: objects, attributes, and relationships. objects were assessed for accuracy and completeness in detection, attributes were evaluated for their relevance and precision in describing the detected objects, and relationships were analyzed for their contextual appropriateness and validity in representing object connections within the scene. Each component was rated on a scale of 1 to 5, where 1 is the lowest and 5 is the highest. The scoring guidelines were as follows: a score of 5 indicated perfection with no corrections required, 4 signified minor issues requiring some tweaking, 3 corresponded to major issues necessitating further improvement, 2 reflected significant issues requiring major revisions, and 1 represented an invalid component deemed beyond repair.

Table 1
Visual Scene Graph Validation by Human Annotators and Weighted Kappa Score

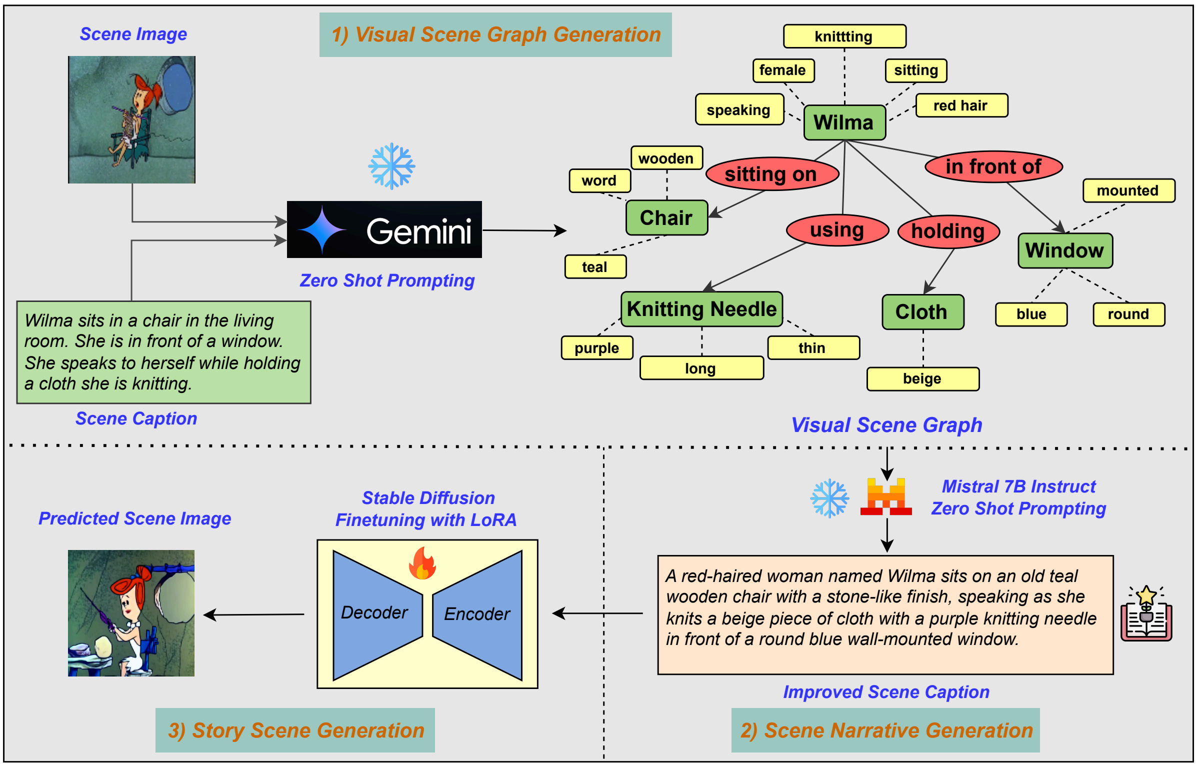
Components	Rating	κ	Agreement
Objects	4.68	0.45	Moderate
Attributes	4.62	0.31	Fair
Relationships	4.41	0.26	Fair

The average ratings for each component, as summarized in Table 1, indicate that object detection received the highest scores, followed by attribute evaluation, while relationships scored the lowest. The relatively lower scores for relationships underscore the challenges in accurately reasoning about inter-object connections within the scene, reflecting the increased complexity of this task. Additionally, we present the Weighted Cohen's Kappa score to evaluate inter-annotator agreement on the VSG validation. We report the average Kappa score calculated across all possible annotator pairs. The results revealed a moderate level of agreement for object detection, while attributes and relationships exhibited fair agreement, with scores following the same trend as the average ratings with objects have higher Kappa score, followed by attributes and then relationships due to increasing complexity in detection.

In summary, the high-quality generated VSGs significantly enhance the original scene captions by providing a more detailed and nuanced representation of the story scenes. These findings highlight the potential of VSGs to improve scene understanding and reasoning in narrative contexts.

4. Methodology

This section outlines our three-stage methodology for improving the story scene narratives. The subsequent sub-sections provide detailed descriptions of each stage. Section 4.1 explains the process of VSG generation from scene images and descriptions using Gemini multi-modal model. Section 4.2 discusses narrative scene generation using factual knowledge derived from the generated VSG. Finally, Section 4.3 presents the fine-tuning of diffusion models for the text-to-story generation task.



Wilma sits in a chair in the living room. She is in front of a window. She speaks to herself while holding a cloth she is knitting.

A red-haired woman named Wilma sits on an old teal wooden chair with a stone-like finish, speaking as she knits a beige piece of cloth with a purple knitting needle in front of a round blue wall-mounted window.

Figure 2: Three-Stage Methodology of Flintstones++. 1) **Visual Scene Graph Generation**: In this stage, the story Scene Image and Scene Caption are fed into the Gemini-Flash model using a structured prompt to generate the Visual Scene Graph. where green boxes denotes objects, yellow boxes denotes attributes and red circles denotes the relationship between the objects. 2) **Scene Narrative Generation**: The Visual Scene Graph is utilized to generate an Improved scene Caption using zero-shot prompting on the Mistral-7B LLM. 3) **Story Scene Generation**: In the final stage, stable diffusion models are fine-tuned with the parameter-efficient LoRA technique to generate the Scene Image from the Improved Scene Caption.

4.1. Visual Scene Graph (VSG) Generation

To extract factual information in the form of a Visual Scene Graph of story scene images, as described in the Stage 1 of Figure 2, we utilize the generative capabilities of the pre-trained Large Vision-Language Model (LVLM) Gemini-Flash [31]. This model generates a Visual Scene Graph by processing the scene image along with its corresponding caption from the original FlintstonesSV dataset. Let $\mathcal{I}$ denote the set of input scene images and $\mathcal{C}$ denote the set of corresponding captions, where each caption $C \in \mathcal{C}$ is paired with an image $I \in \mathcal{I}$. Let $\mathcal{M}$ represent Gemini-Flash as the pre-trained LVLM, and $\mathcal{P}(I, C)$ indicate the handcrafted zero shot prompt [1] formed by combining an image I with its corresponding caption C. For each pair (I, C) from the dataset, the visual scene graph is generated as follows:

$$\mathcal{G}_I = \mathcal{M}(\mathcal{P}(I, C)),$$

where $\mathcal{G}_I = (\mathcal{O}_I, \mathcal{A}_I, \mathcal{R}_I)$ represents the visual scene graph for image I. Here, $\mathcal{O}_I$ denotes the set of detected objects, $\mathcal{A}_I = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ represents the attributes associated with each object (α_i for o_i), and $\mathcal{R}_I = \{\rho_{ij}\}$ captures the relationships between objects o_i and o_j.

4.2. Scene Narrative Generation

To generate enriched scene narratives for the images in the FlintstonesSV dataset, as described in the Stage 2 of Figure 2, For each image I, the corresponding VSG $\mathcal{G}_I$ extracted in Stage 1 is then converted into a natural language prompt $\mathcal{P} = \mathcal{F}(\mathcal{G}_I)$. Where $\mathcal{F}(\cdot)$ is a formatting function which structures the VSG into an instruction that can be effectively processed by a large language model (LLM).

The prompt $\mathcal{P}$ is then processed by the Mistral-7B Instruct [32] model $\mathcal{M}$ to generate the enriched scene narrative $\mathcal{E} = \mathcal{M}(\mathcal{P})$. These enriched narratives $\mathcal{E}$ enhance the original dataset by incorporating additional details including backgrounds, character attributes, precise character positioning, inter-object relationships, and the presence of other high-level objects. As a result, the enriched narratives provide a more comprehensive context, improving the dataset's utility for downstream tasks like story visualization.

4.3. Story Scene Generation

For story scene generation from narrative prompts, as we can see in the Stage 3 of Figure 2, we leverage open-source pre-trained text-to-image diffusion models to generate images corresponding to the provided story scene narratives. To tailor the model to our specific dataset, we fine-tune the model using the LoRA [33] parameter-efficient fine-tuning technique. Let $\mathcal{M}$ represent the pre-trained text-to-image diffusion model. The model takes as input a prompt $\mathcal{P}$, which represents the story scene narrative, and outputs a generated image $\mathcal{I}_{\text{pred}}$, where:

$$\mathcal{I}_{\text{pred}} = \mathcal{M}(\mathcal{P}).$$

During fine-tuning, the model $\mathcal{M}$ is optimized to minimize the Stable Diffusion loss $\mathcal{L}$, which is computed based on the denoising process within the diffusion framework. Specifically, for a latent representation $\mathcal{Z}_0$, a time step t, and noise ϵ, the Stable Diffusion loss is defined as:

$$\mathcal{L} = \mathbb{E}_{\mathcal{Z}_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathcal{Z}_t, t)\|^2\right].$$

In this equation, $\mathcal{Z}_t$ refers to the noisy latent variable at time t, and ϵ denotes the noise added during the forward process. The term $\epsilon_\theta(\mathcal{Z}_t, t)$ represents the predicted noise at time t, estimated by the model $\mathcal{M}$, which is parameterized by θ, and $\|\cdot\|^2$ indicates the squared Euclidean norm.

Once the model is fine-tuned, it is loaded to be tested on unseen story narratives to generate the corresponding visual scenes, demonstrating its capability to produce relevant story scenes.

5. Experiments and Results

5.1. Experimental Setup

We employed three open-source pre-trained text-to-image models—SDXL Base-1.0 [2], CompVis Stable Diffusion V4, and Stable Diffusion-2 [34]—to evaluate the effectiveness of the FlintstonesSV++ dataset. These models were fine-tuned using parameter-efficient methods, specifically the Low-Rank Adaptation (LoRA) technique [33], for the scene generation task based on scene narratives. During fine-tuning, we used a batch size of 8, trained the models for 10 epochs, and applied a cosine learning rate scheduler. All other hyper-parameters were kept at their default values to ensure consistency in the evaluation process.

[1] Prompts for VSG Extraction and Story Narrative Generation are given in the GitHub Repository - FlintstonesSV++

5.2. Evaluation Metrics

To evaluate the effectiveness of visual scene generation on the improved Flintstones++ dataset compared to the original Flintstones dataset, we used two metrics: **FID Score** [35] and **CLIP Score** [36]. The **FID (Fréchet Inception Distance)** measures the quality of generated images by comparing their feature distributions with those of real images, using the Inception-V3 model. A lower FID score indicates better image quality. The **CLIP Score** assesses how well the generated scene aligns with the story narrative. It is computed by passing the generated image and the story narrative through the pre-trained OpenAI CLIP model's image and text encoders, respectively. The embeddings are compared using cosine similarity, with a higher CLIP score indicating better alignment. We report the average FID and CLIP scores over all samples on the test set.

5.3. Quantitative Results

This section presents the fine-tuning results on the FlintstonesSV++ and FlintstonesSV datasets using three different text-to-image generation models. As shown in Table 2, the CLIP score improves by 6.23%, 4.85%, and 4.78% across the three models, indicating a better alignment between the scene captions and the generated images. On average, a 5.2% improvement in the CLIP score demonstrates the effectiveness of the FlintstonesSV++ captions in enhancing downstream tasks. These captions enable the generated scene images to align more accurately with their corresponding captions. Moreover, the FID score decreases by 14%, 3%, and approximately 1% for the text-to-image generation task when using FlintstonesSV++ compared to FlintstonesSV. While some models show only minor improvements in the average FID score across the full test set due to their pre-training capabilities, still their alignment between generated images and scene narratives also improves significantly with FlintstonesSV++. These enhancements in alignment and image generation make FlintstonesSV++ a superior choice over FlintstonesSV for downstream tasks such as story visualization and story continuation.

FlintstonesSV++ offers detailed scene narratives that capture not only the spatial positions of characters and objects but also the relationships between them, which are crucial for accurate scene transitions. These improvements address the limitations of FlintstonesSV, as evidenced by the quantitative results.

Table 2
Text-to-Story Generation Results on Different Text-to-Image Generation Models

Dataset	SDXL Base 1.0		Stable Diffusion V4		Stable Diffusion 2	
	CLIP (↑)	FID (↓)	CLIP (↑)	FID (↓)	CLIP (↑)	FID (↓)
Flintstones	0.2727	77.72	0.2841	52.02	0.2958	42.18
Flintstones++ (Ours)	**0.3350**	**63.36**	**0.3326**	**49.87**	**0.3436**	**41.52**

5.4. Qualitative Results Analysis

To evaluate the impact of the FlintstonesSV++ dataset with enhanced scene captions, we present the results of scene images generated by fine-tuned Stable Diffusion models on a text-to-image generation task. The model was trained to generate story scenes based on input narratives. As shown in Figure 3, we compare three samples, each with its ground truth image and the predicted image generated using scene descriptions from both the Flintstone and FlintstonesSV++ datasets. The Flintstone dataset results reveal several limitations. In example **(a)**, duplicate characters appear, and while multiple windows are present, curtains are missing. In example **(b)**, the background home and the characters' emotional expressions differ from the ground truth. In example **(c)**, Betty and Wilma are not depicted on the beach, and the umbrella is missing. The FlintstonesSV++ dataset addresses these shortcomings by using VSG enhanced captions and factual information added by visual scene graph is highlighted in blue , which provide more accurate and detailed cues. As a result, the generated scenes align better with the input narratives. These improvements demonstrate the effectiveness of the FlintstonesSV++ dataset in enhancing both story visualization and story-continuation tasks.

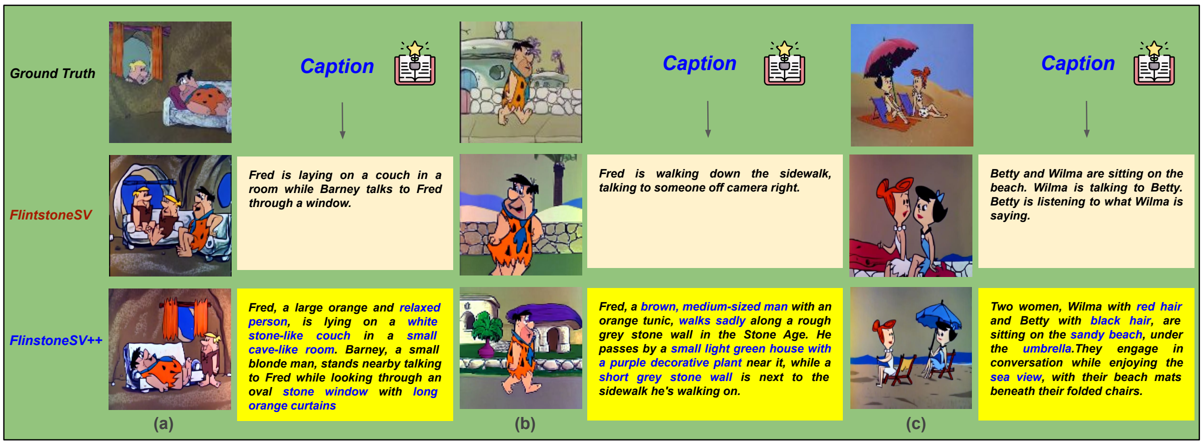

Ground Truth
FlintstoneSV
FlinstoneSV++
Caption
Caption
Caption
Fred is laying on a couch in a room while Barney talks to Fred through a window.
Fred, a large orange and relaxed person, is lying on a white stone-like couch in a small cave-like room. Barney, a small blonde man, stands nearby talking to Fred while looking through an oval stone window with long orange curtains
(a)
Fred is walking down the sidewalk, talking to someone off camera right.
Fred, a brown, medium-sized man with an orange tunic, walks sadly along a rough grey stone wall in the Stone Age. He passes by a small light green house with a purple decorative plant near it, while a short grey stone wall is next to the sidewalk he's walking on.
(b)
Betty and Wilma are sitting on the beach. Wilma is talking to Betty. Betty is listening to what Wilma is saying.
Two women, Wilma with red hair and Betty with black hair, are sitting on the sandy beach, under the umbrella.They engage in conversation while enjoying the sea view, with their beach mats beneath their folded chairs.
(c)

Figure 3: Comparison of ground truth scene images with generated scene images produced by fine-tuned diffusion models for the scene narrative-to-image generation task on the FlintstonesSV and FlintstonesSV++ datasets. blue color highlighted text denotes the factual information added by our visual scene graph.

6. Conclusion

In this paper, we introduced FlintstonesSV++, an improved version of the FlintstonesSV dataset, enhanced using Visual Scene Graphs (VSG) and Large Language Models (LLMs). Our approach enriches the dataset by incorporating factual information that was previously absent, making it more suitable for benchmarking downstream tasks such as story visualization and story continuation. Experimental results demonstrate that FlintstonesSV++ achieves superior performance in text-to-story generation task, highlighting the effectiveness of our enhancements. These findings establish FlintstonesSV++ as successor to FlintstonesSV, offering richer and more detailed scene descriptions through VSG and LLMs, thereby improving its utility in narrative-based AI applications.

7. Limitations

Our dataset construction leverages pre-trained models, specifically the Gemini-Flash Large Vision-Language Model for extracting visual scene graphs from images and the Mistral LLM for generating scene stories based on these graphs. The accuracy of the extracted scene graphs and generated stories is inherently constrained by the pre-trained capabilities of these models. Since these models are utilized in a zero-shot prompting setting, their outputs may exhibit biases present in their training data. Despite these limitations, our experimental results demonstrate significant improvements over the original dataset. Additionally, we experiment with the diffusion models for per-frame story scene visualization in this paper to showcase the effectiveness of scene narrative.

Acknowledgement

We thank the anonymous reviewers for their insights on this work. This publication has emanated from research conducted with the financial support of Research Ireland under Grant Number SFI/12/RC/2289_P2 (Insight), co-funded by the European Regional Development Fund.

References

[1] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al., Improving image generation with better captions, Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf 2 (2023) 8.

[2] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, Sdxl: Improving latent diffusion models for high-resolution image synthesis, arXiv preprint arXiv:2307.01952 (2023).

[3] J. Baldridge, J. Bauer, M. Bhutani, N. Brichtova, A. Bunner, K. Chan, Y. Chen, S. Dieleman, Y. Du, Z. Eaton-Rosen, et al., Imagen 3, arXiv preprint arXiv:2408.07009 (2024).

[4] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al., Scaling rectified flow transformers for high-resolution image synthesis, in: Forty-first International Conference on Machine Learning, 2024.

[5] Y. Qin, Z. Shi, J. Yu, X. Wang, E. Zhou, L. Li, Z. Yin, X. Liu, L. Sheng, J. Shao, et al., Worldsimbench: Towards video generation models as world simulators, arXiv preprint arXiv:2410.18072 (2024).

[6] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al., Stable video diffusion: Scaling latent video diffusion models to large datasets, arXiv preprint arXiv:2311.15127 (2023).

[7] A. Sharma, A. Yu, A. Razavi, A. Toor, A. Pierson, A. Gupta, A. Waters, A. van den Oord, D. Tanis, D. Erhan, E. Lau, E. Shaw, G. Barth-Maron, G. Shaw, H. Zhang, H. Nandwani, H. Moraldo, H. Kim, I. Blok, J. Bauer, J. Donahue, J. Chung, K. Mathewson, K. David, L. Espeholt, M. van Zee, M. McGill, M. Narasimhan, M. Wang, M. Bińkowski, M. Babaeizadeh, M. T. Saffar, N. de Freitas, N. Pezzotti, P.-J. Kindermans, P. Rane, R. Hornung, R. Riachi, R. Villegas, R. Qian, S. Dieleman, S. Zhang, S. Cabi, S. Luo, S. Fruchter, S. Nørly, S. Srinivasan, T. Pfaff, T. Hume, V. Verma, W. Hua, W. Zhu, X. Yan, X. Wang, Y. Kim, Y. Du, Y. Chen, Veo (2024). URL: https://deepmind.google/technologies/veo/.

[8] C. Liu, H. Wu, Y. Zhong, X. Zhang, Y. Wang, W. Xie, Intelligent grimm - open-ended visual storytelling via latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6190–6200.

[9] T. Rahman, H.-Y. Lee, J. Ren, S. Tulyakov, S. Mahajan, L. Sigal, Make-a-story: Visual memory conditioned consistent story generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2493–2502.

[10] X. Pan, P. Qin, Y. Li, H. Xue, W. Chen, Synthesizing coherent story with auto-regressive latent diffusion models, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2920–2930.

[11] S. Zheng, Y. Fu, Temporalstory: Enhancing consistency in story visualization using spatial-temporal attention, arXiv e-prints (2024) arXiv–2407.

[12] T. Gupta, D. Schwenk, A. Farhadi, D. Hoiem, A. Kembhavi, Imagine this! scripts to compositions to videos, 2018. URL: https://arxiv.org/abs/1804.03608. arXiv:1804.03608.

[13] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, J. Gao, Storygan: A sequential conditional gan for story visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International journal of computer vision 123 (2017) 32–73.

[15] B. Li, Word-level fine-grained story visualization, in: European Conference on Computer Vision, Springer, 2022, pp. 347–362.

[16] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, J. Gao, Storygan: A sequential conditional gan for story visualization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6329–6338.

[17] A. Maharana, M. Bansal, Integrating visuospatial, linguistic, and commonsense structure into story visualization, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6772–6786. URL: https://aclanthology.org/2021.emnlp-main.543/. doi:10.18653/v1/2021.emnlp-main.543.

[18] A. Maharana, D. Hannan, M. Bansal, Improving generation and evaluation of visual stories via semantic consistency, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2427–2442. URL: https://aclanthology.org/2021.naacl-main.194/. doi:10.18653/v1/2021.naacl-main.194.

[19] Y.-Z. Song, Z. Rui Tam, H.-J. Chen, H.-H. Lu, H.-H. Shuai, Character-preserving coherent story visualization, in: European Conference on Computer Vision, Springer, 2020, pp. 18–33.

[20] A. Maharana, D. Hannan, M. Bansal, Storydall-e: Adapting pretrained text-to-image transformers for story continuation, in: European Conference on Computer Vision, Springer, 2022, pp. 70–87.

[21] X. Shen, M. Elhoseiny, Large language models as consistent story visualizers, arXiv preprint arXiv:2312.02252 (2023).

[22] W. Wang, C. Zhao, H. Chen, Z. Chen, K. Zheng, C. Shen, Autostory: Generating diverse storytelling images with minimal human efforts, International Journal of Computer Vision (2024) 1–22.

[23] T. Qian, J. Chen, S. Chen, B. Wu, Y.-G. Jiang, Scene graph refinement network for visual question answering, IEEE Transactions on Multimedia 25 (2022) 3950–3961.

[24] V. Damodaran, S. Chakravarthy, A. Kumar, A. Umapathy, T. Mitamura, Y. Nakashima, N. Garcia, C. Chu, Understanding the role of scene graphs in visual question answering, arXiv preprint arXiv:2101.05479 (2021).

[25] M. Hildebrandt, H. Li, R. Koner, V. Tresp, S. Günnemann, Scene graph reasoning for visual question answering, arXiv preprint arXiv:2007.01072 (2020).

[26] J. Shi, H. Zhang, J. Li, Explainable and explicit visual reasoning over scene graphs, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8376–8384.

[27] H. Tian, N. Xu, A.-A. Liu, C. Yan, Z. Mao, Q. Zhang, Y. Zhang, Mask and predict: Multi-step reasoning for scene graph generation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4128–4136.

[28] Z. Wang, H. You, L. H. Li, A. Zareian, S. Park, Y. Liang, K.-W. Chang, S.-F. Chang, Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning, in: Proceedings of the AAAI conference on artificial intelligence, volume 36, 2022, pp. 5914–5922.

[29] X. Li, S. Jiang, Know more say less: Image captioning based on scene graphs, IEEE Transactions on Multimedia 21 (2019) 2117–2130.

[30] Y. Zhong, L. Wang, J. Chen, D. Yu, Y. Li, Comprehensive image captioning via scene graph decomposition, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, 2020, pp. 211–229.

[31] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al., Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, arXiv preprint arXiv:2403.05530 (2024).

[32] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.