FlintstonesSV++ : Improving Story Narration using Visual Scene Graph

Taighde Éireann
Research Ireland

ECIR
2025

Insight
RESEARCH IRELAND CENTRE FOR DATA ANALYTICS

Janak Kapuriya, Paul Buitelaar

University of Galway, Ireland

HOST INSTITUTIONS

DCU
Ollscoil Chathair Bhaile Átha Cliath
Dublin City University

Ollscoil na Gaillimhe
University of Galway

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

PARTNER INSTITUTIONS

Maynooth University
National University of Ireland Maynooth

Tyndall
National Institute

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

University of Limerick
Ollscoil Luimnigh

FUNDED BY

Taighde Éireann
Research Ireland

Story Visualization (SV)

- Given a story narrative, generate a sequence of scene images that convey the meaning of the narratives

Story Visualization Example

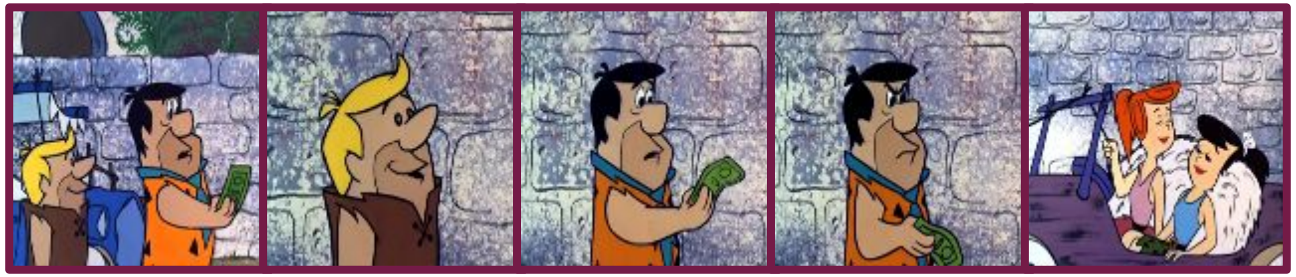
- Given a story narrative, generate a sequence of scene images that convey the meaning of the narratives

Scene 1 - Fred and Barney are outside, standing next to a car. Fred holding money in his hand while speaking to someone.
Scene 2 - Barney is outside pointing at something. While he is pointing he is saying something.
Scene 3 - Fred is holding money in the room.
Scene 4 - Fred looks at some money and talks in a store.
Scene 5 - Betty and Wilma are sitting in a car. Wilma tugs at a rope while Betty leans back in her seat.

Narratives

Story Visualization Example

- Given a story narrative, generate a sequence of scene images that convey the meaning of the narratives



Scene 1 **Scene 2** **Scene 3** **Scene 4** **Scene 5**

Scene 1 - Fred and Barney are outside, standing next to a car. Fred holding money in his hand while speaking to someone.
Scene 2 - Barney is outside pointing at something. While he is pointing he is saying something.
Scene 3 - Fred is holding money in the room.
Scene 4 - Fred looks at some money and talks in a store.
Scene 5 - Betty and Wilma are sitting in a car. Wilma tugs at a rope while Betty leans back in her seat.

Narratives

Story Continuation (SC)

- Given an initial frame and story narratives, generate a sequence of coherent scene images that extend the initial frame based on the progression of the narratives

Story Continuation Example

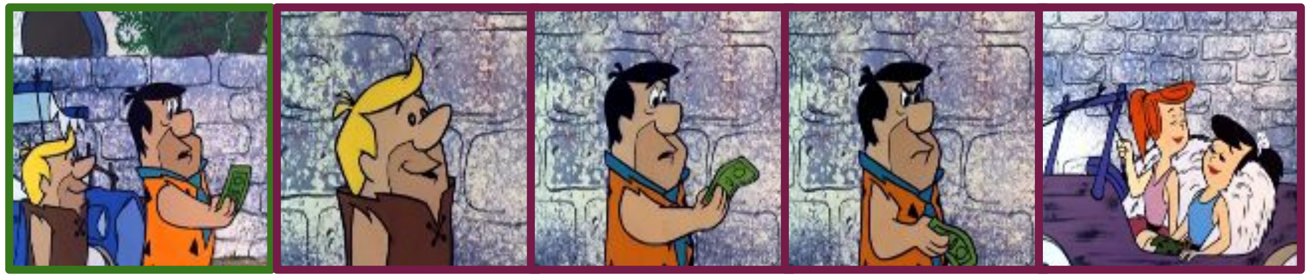
- Given an initial frame and story narratives, generate a sequence of coherent scene images that extend the initial frame based on the textual progression of the narratives

Scene 1 - Fred and Barney are outside, standing next to a car. Fred holding money in his hand while speaking to someone.
Scene 2 - Barney is outside pointing at something. While he is pointing he is saying something.
Scene 3 - Fred is holding money in the room.
Scene 4 - Fred looks at some money and talks in a store.
Scene 5 - Betty and Wilma are sitting in a car. Wilma tugs at a rope while Betty leans back in her seat.

Narratives

Story Continuation Example

- Given an initial frame and story narratives, generate a sequence of coherent scene images that extend the initial frame based on the textual progression of the narratives



Scene 1 - Fred and Barney are outside, standing next to a car. Fred holding money in his hand while speaking to someone.
Scene 2 - Barney is outside pointing at something. While he is pointing he is saying something.
Scene 3 - Fred is holding money in the room.
Scene 4 - Fred looks at some money and talks in a store.
Scene 5 - Betty and Wilma are sitting in a car. Wilma tugs at a rope while Betty leans back in her seat.

Narratives

Related Works

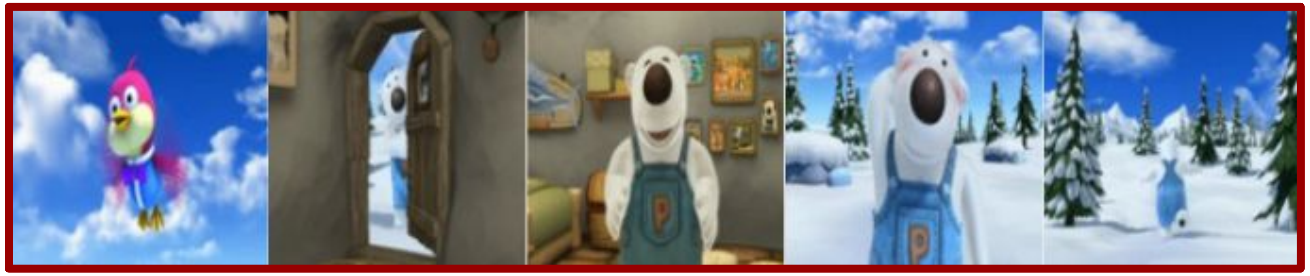
- **GAN based Story Visualization and Continuation**
  - StoryGAN (Y, Li et al. 2019)
  - CP-CSV (Song, Y.Z et al. 2020)
  - DUCO-GAN (Maharana, A et al. 2021)
  - VLC (Maharana, A et al. 2021)
- **Diffusion based Story Visualization and Continuation**
  - Make-A-story (Rahman, Tanzila et al. 2023)
  - StoryGPT-V (Shen, Xiaoqian et al. 2023)
  - ARLDM (X. Pan et al. 2024)
  - TemporalStory (Zheng, Sixiao et al. 2024)

Benchmark Datasets

- **FlintstonesSV** (Tanmoy, Gupta et al. 2018)
  - 7 main characters
  - Train, Val, Test (20132, 2071, 2309)
- **PororoSV** (Yitong, Li et al. 2018)
  - 8 main characters
  - Train, Val, Test (10191, 2334, 2208)
- Each sample consists of 5 pairs of (scene image, narrative)
- Used for story visualization and story continuation benchmarking

Benchmark Datasets

FlintstonesSV



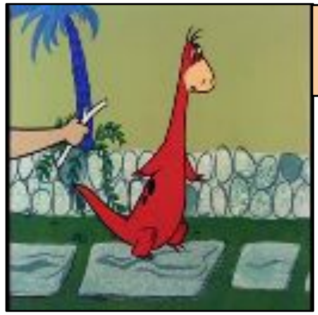
PororoSV

Limitations of FlintstonesSV

- **Scene narrative described only**
  - *character's name*
  - *activity*
  - *location*
- **Missing Important Details**
  - *character attributes*
  - *precise character position in scene*
  - *detailed background Information*
  - *high level objects*
  - *relationship of objects with other objects and characters in scene*

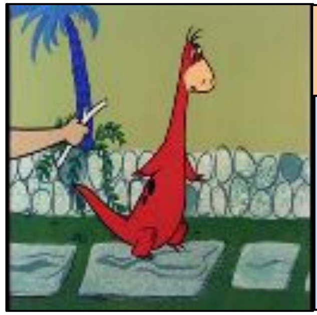
Limitation Examples

Limitation Examples



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

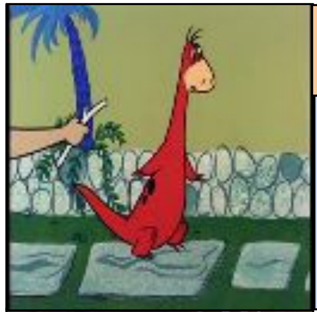
Limitation Examples



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

- missing background information
- missing precise position of dino in scene

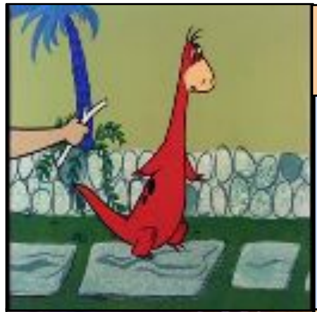
Limitation Examples



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

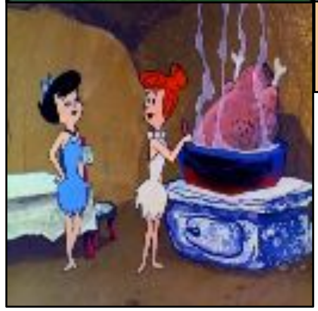
- missing background information
- missing precise position of dino in scene

Limitation Examples



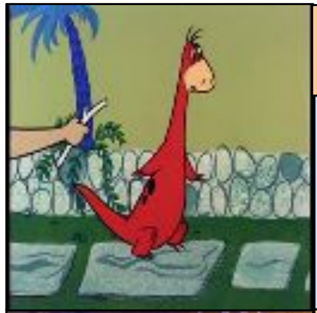
Red color dino is in the yard looking at a stick.

- missing background information
- missing precise position of dino in scene



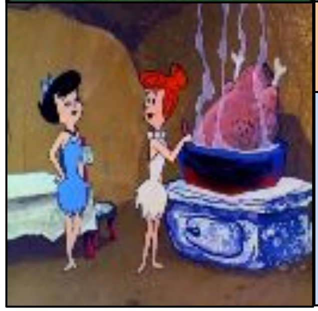
Betty and Wilma are in the kitchen. Betty is talking to Wilma. Wilma is cooking

Limitation Examples



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

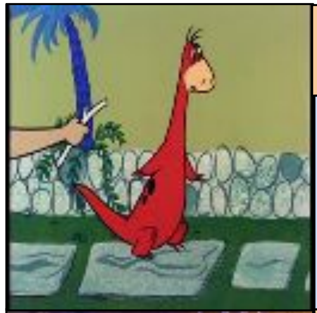
- missing background information
- missing precise position of dino in scene



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

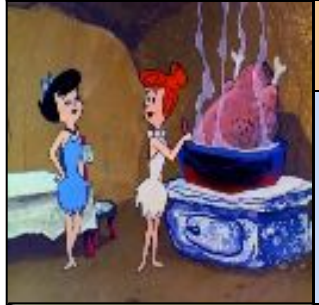
- missing type of food being cook
- missing utensil used for cooking and its color

Limitation Examples



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

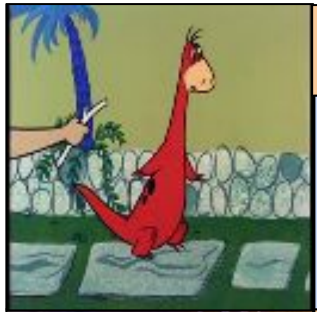
- missing background information
- missing precise position of dino in scene



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

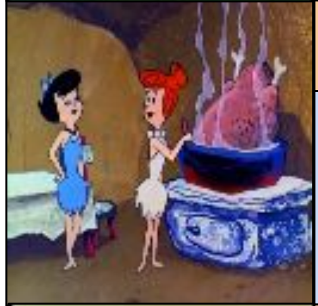
- missing type of food being cook
- missing utensil used for cooking and its color

Limitation Examples



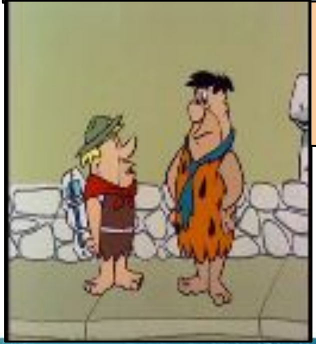
Red color dino is in the yard looking at a stick.

- missing background information
- missing precise position of dino in scene



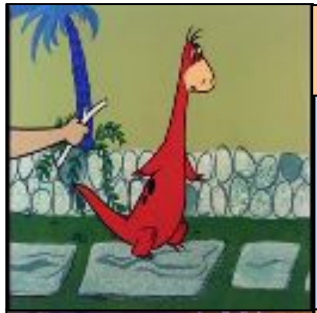
Betty and Wilma are in the kitchen. Betty is talking to Wilma. Wilma is cooking

- missing type of food being cook
- missing utensil used for cooking and its color



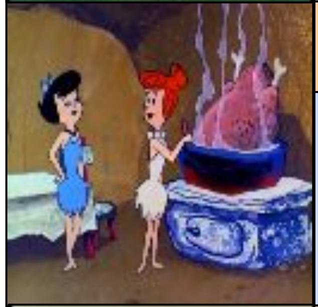
Fred and Barney are standing on a sidewalk. Barney is speaking to Fred, while Fred listens silently with his hands on his hips.

Limitation Examples



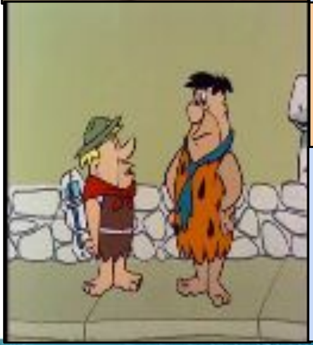
Red color dino is in the yard looking at a stick.

- missing background information
- missing precise position of dino in scene



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

- missing type of food being cook
- missing utensil used for cooking and its color



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

- missing details about character apparel
- missing background elements like the wall
- missing spatial position of characters in scene

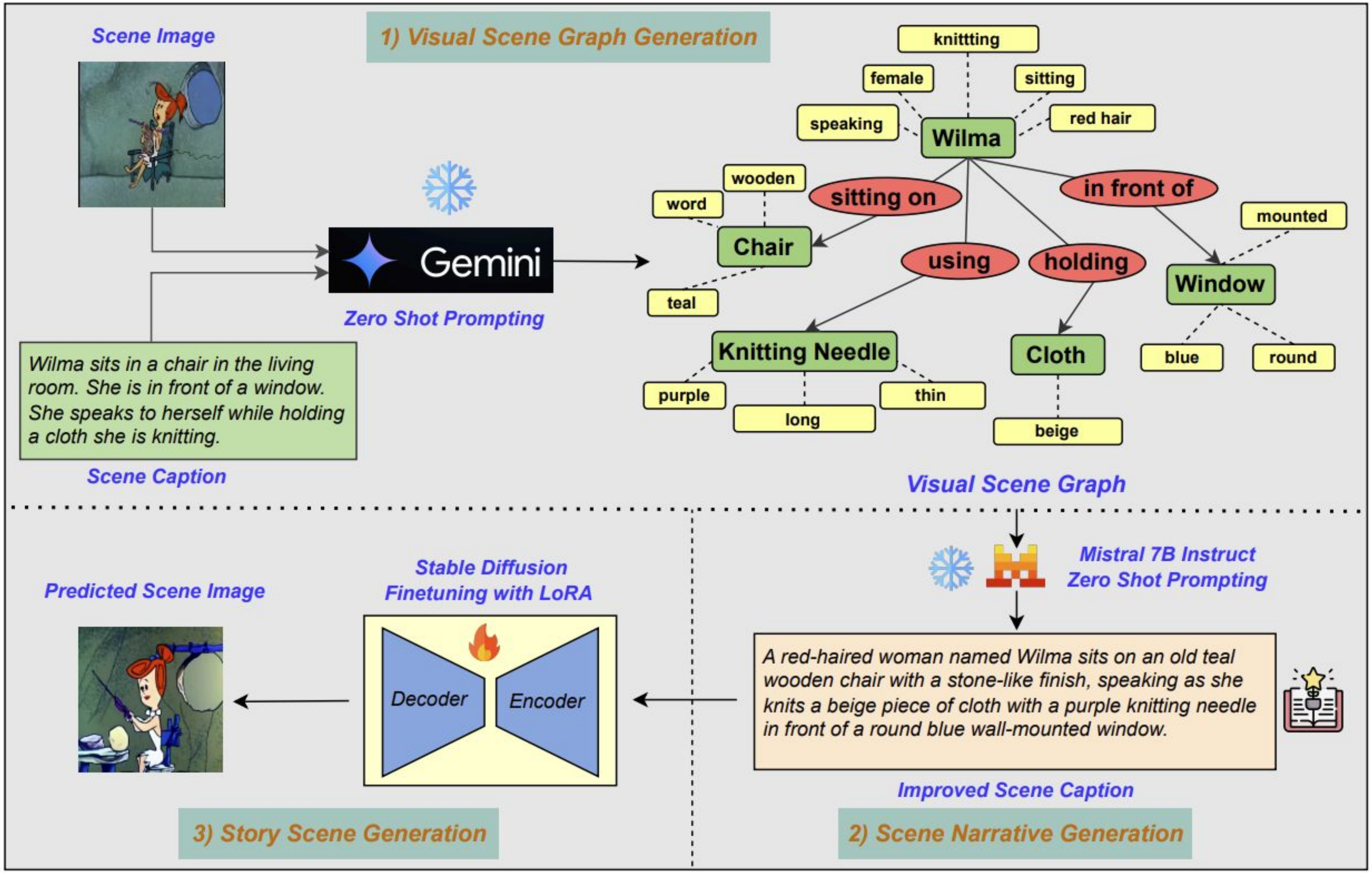
Limitations of FlintstonesSV dataset

- Missing factual details in scene narrative
- these gaps limit the dataset’s ability to capture the complete essence of a story scene.
- models trained on this benchmark dataset often struggle with generating or continuing stories that are contextually rich and detailed
- *these findings highlight the need for improved scene narratives to enhance the performance of narrative-based AI applications.*

Visual Scene Graph (VSG) *(R. Krishna et al. 2016)*

- VSG represents *factual details* from images in the form of Objects, Attributes and Relationships
- **VSG Related Works**
  - *visual question answering (V. Damodaran et al. 2021, T. Qian et al. 2022)*
  - *image captioning (X. Li et al. 2019, Y. Zhong et al. 2020)*
  - *visual scene reasoning (H. Tian et al. 2021, Z. Wang et al. 2022)*
- We utilize Visual Scene Graphs to add factual details from scene images to enhance the scene narrative of the FlinstonesSV dataset

FlintstonesSV++ Methodology



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.

Visual Scene Graph (VSG) Human Evaluation

- Components of the Visual Scene Graph
  - Objects: The entities present in the scene.
  - Attributes: Descriptive features associated with each object.
  - Relationships: The interactions or spatial connections between the objects.
- Human Evaluation done by 7 annotators on 10 random VSG samples

Rating Guidelines

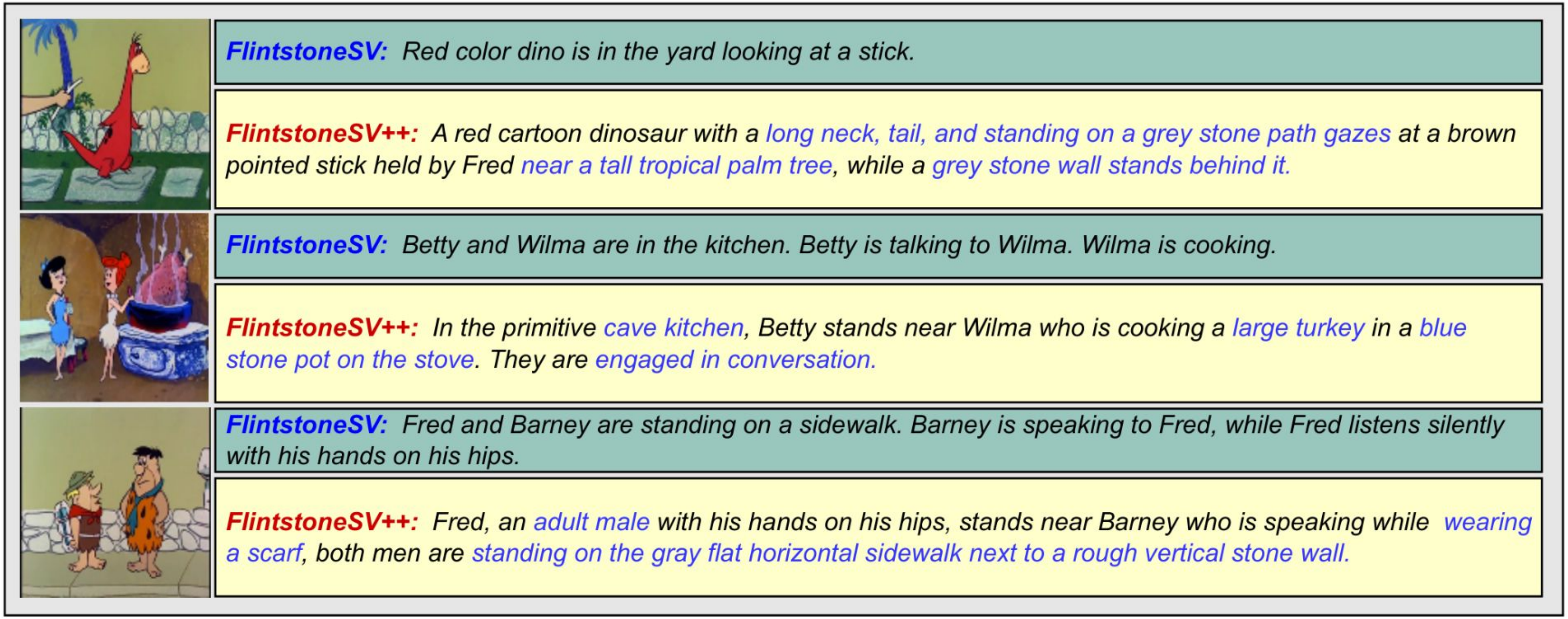
- **Ordinal Rating**
  - 1 to 5
  - 1 lowest
  - 5 highest

- **Rating Description**
  - 5 - Perfect *(No Correction Required)*
  - 4 - Minor Issues *(Some Tweakings Required)*
  - 3 - Major Issues *(Need Further Improvement)*
  - 2 - Significant Issues *(Need Major Improvement)*
  - 1 - Rubbish *(Beyond Repair)*

VSG Human Evaluation Results

Components	Avg. Rating	Cohen's Kappa	Agreement
Objects	4.68	0.45	Moderate
Attributes	4.62	0.31	Fair
Relationships	4.41	0.26	Fair

Effectiveness of FlintstonesSV++



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 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.

Story Visualization Experiments

- **Diffusion Models**
  - *SDXL Base 1.0*
  - *Stable Diffusion V4*
  - *Stable Diffusion 2*
- **Hyperparameters**
  - *10 epoch*
  - *8 batch size*
  - *cosine scheduler*
  - *other params were kept default*

Evaluation Metrics

- FID Score *(Fréchet Inception Distance)*
  - Measures the quality of generated images by comparing feature distributions of generated image with real images
  - lower is better

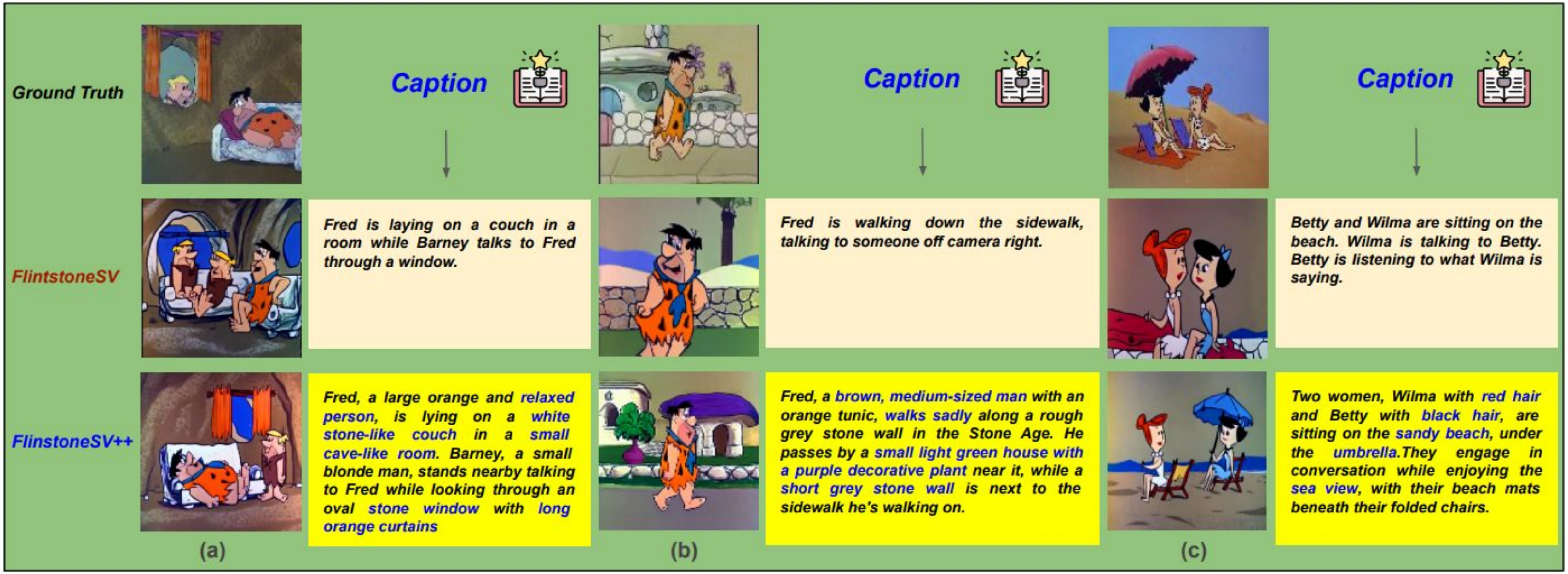
- CLIP Score
  - Assesses alignment between generated scene and story narrative
  - higher is better

Results

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

- average *5.20%* boost in alignment score
- average *5.72%* boost in image generation quality

Qualitative Results



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

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.

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.

Conclusion

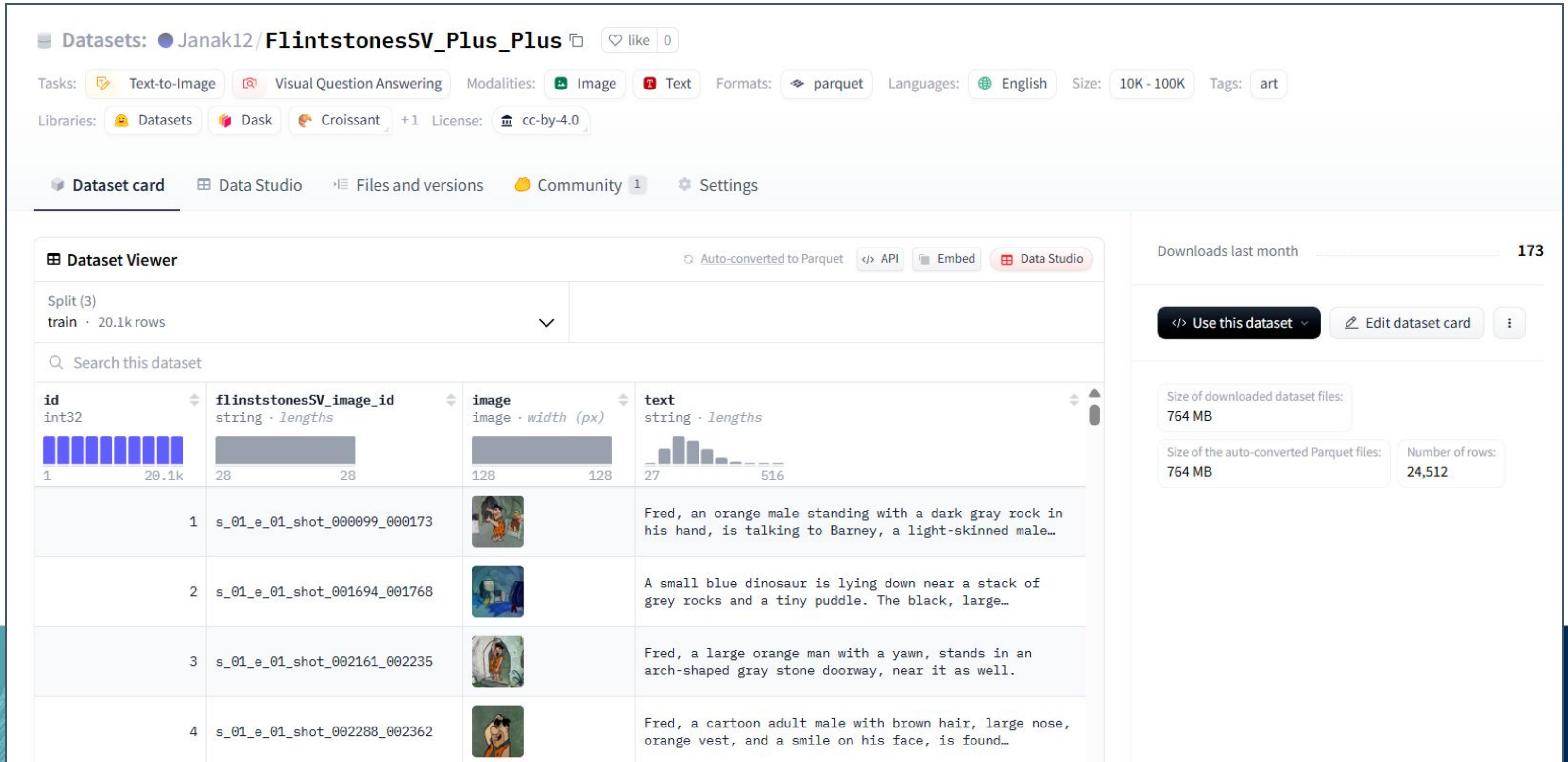
- Visual Scene Graph adds required factual information which is crucial for complete scene understanding for a task like story visualization and story continuation
- FlintstonesSV++ achieves superior performance compared to FlintstonesSV for the story narrative to scene generation task
- FlintstonesSV++ demonstrates rich and detailed scene narratives, which provide a resource for narrative-based AI applications

Open Source

- **Dataset** - ***FlintstonesSV_Plus_Plus*** 🤗🤗🤗

Github

Paper and Code



Datasets: Janak12/FlintstonesSV_Plus_Plus ♡ like 0

Tasks: Text-to-Image · Visual Question Answering · Modalities: Image · Text · Formats: parquet · Languages: English · Size: 10K - 100K · Tags: art

Libraries: Datasets · Dask · Croissant · +1 · License: cc-by-4.0

Dataset Viewer

Split (3)
train · 20.1k rows

Search this dataset

id int32 1 · 20.1k	flinststonesSV_image_id string · lengths 28 · 28	image image · width (px) 128 · 128	text string · lengths 27 · 516
1	s_01_e_01_shot_000099_000173		Fred, an orange male standing with a dark gray rock in his hand, is talking to Barney, a light-skinned male…
2	s_01_e_01_shot_001694_001768		A small blue dinosaur is lying down near a stack of grey rocks and a tiny puddle. The black, large…
3	s_01_e_01_shot_002161_002235		Fred, a large orange man with a yawn, stands in an arch-shaped gray stone doorway, near it as well.
4	s_01_e_01_shot_002288_002362		Fred, a cartoon adult male with brown hair, large nose, orange vest, and a smile on his face, is found…

Downloads last month 173

Size of downloaded dataset files: 764 MB

Size of the auto-converted Parquet files: 764 MB

Number of rows: 24,512

References

- *Rahman, Tanzila, et al. "Make-a-story: Visual memory conditioned consistent story generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.*
- *Shen, Xiaoqian, and Mohamed Elhoseiny. "StoryGPT-V: Large Language Models as Consistent Story Visualizers." (2023).*
- *Tao, Ming, et al. "StoryImager: A Unified and Efficient Framework for Coherent Story Visualization and Completion." arXiv preprint arXiv:2404.05979 (2024).*
- *Zheng, Sixiao, and Yanwei Fu. "TemporalStory: Enhancing Consistency in Story Visualization using Spatial-Temporal Attention." arXiv preprint arXiv:2407.09774 (2024).*
- *Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International journal of computer vision 123 (2017): 32-73.*

References

- *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/CVF Conference on Computer Vision and Pattern Recognition. pp. 6329–6338 (2019)*
- *Maharana, A., Hannan, D., Bansal, M.: Improving generation and evaluation of visual stories via semantic consistency. arXiv preprint arXiv:2105.10026 (2021)*
- *Maharana, A., Bansal, M.: Integrating visuospatial, linguistic and commonsense structure into story visualization. arXiv preprint arXiv:2110.10834 (2021)*
- *Song, Y.Z., Tam, Z.R., Chen, H.J., Lu, H.H., Shuai, H.H.: Character-preserving coherent story visualization. In: European Conference on Computer Vision. pp. 18–33. Springer (2020)*

Thank you

Questions ?