Make-A-Story: Visual Memory Conditioned Consistent Story Generation

Tanzila Rahman[1,3] Hsin-Ying Lee[2] Jian Ren[2] Sergey Tulyakov[2]
Shweta Mahajan[1,3] Leonid Sigal[1,3]

1 UBC 2 3

TUE-AM-238

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Problem Formulation

Imagine if you can just tell story to your children….

Wilma and Betty are in a room. Wilma is talking to Betty and standing on a pedestal while Betty hems her dress. **She** turns to grab quills from a porcupine.

. . .

Wilma is in a room. **She** is speaking to someone while looking over her right shoulder.

. . .

Wilma and Betty are in the living room. Betty talks and then Wilma moves her head.

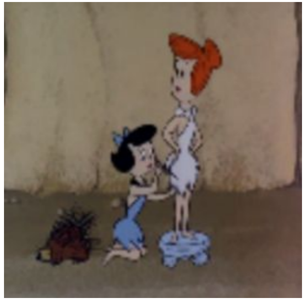
. . .

Wilma is standing in the room talking.

Problem Formulation

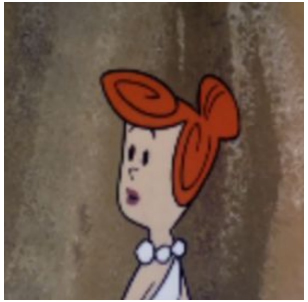
And the story can be generated automatically by maintaining resolution of references and consistency in subject/background appearance:

Wilma and Betty are in a room. Wilma is talking to Betty and standing on a pedestal while Betty hems her dress. **She** turns to grab quills from a porcupine.



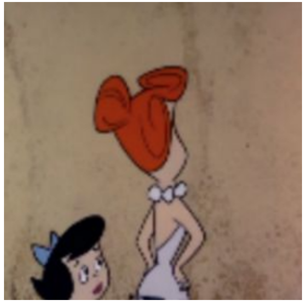
. . .

Wilma is in a room. **She** is speaking to someone while looking over her right shoulder.



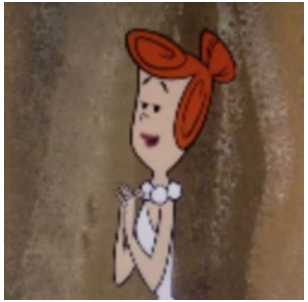
. . .

Wilma and Betty are in the living room. Betty talks and then Wilma moves her head.



. . .

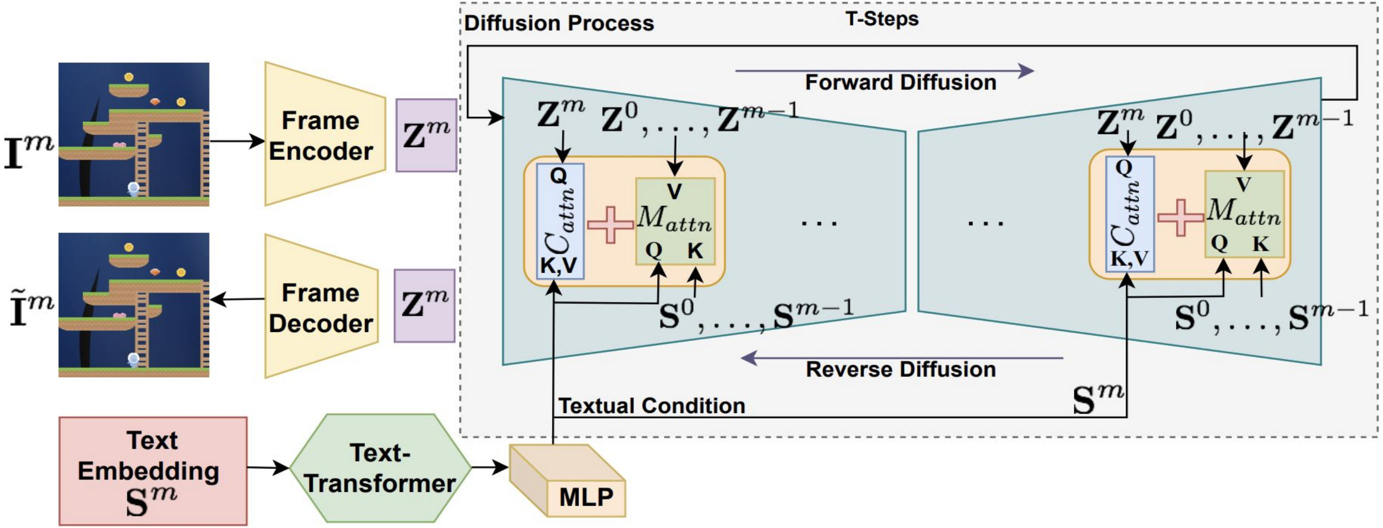
Wilma is standing in the room talking.



Our Contributions

- Story-LDM: A novel autoregressive deep generative framework for the task of story generation.
- A novel memory-attention mechanism to generate consistent stories.
- Extend existing datasets to include more complex scenarios to validate co-reference resolution for character and background consistency in the visual domain.
- Novel evaluation metrics to evaluate for foreground (character) as well as background consistency.

Story-LDM Architecture



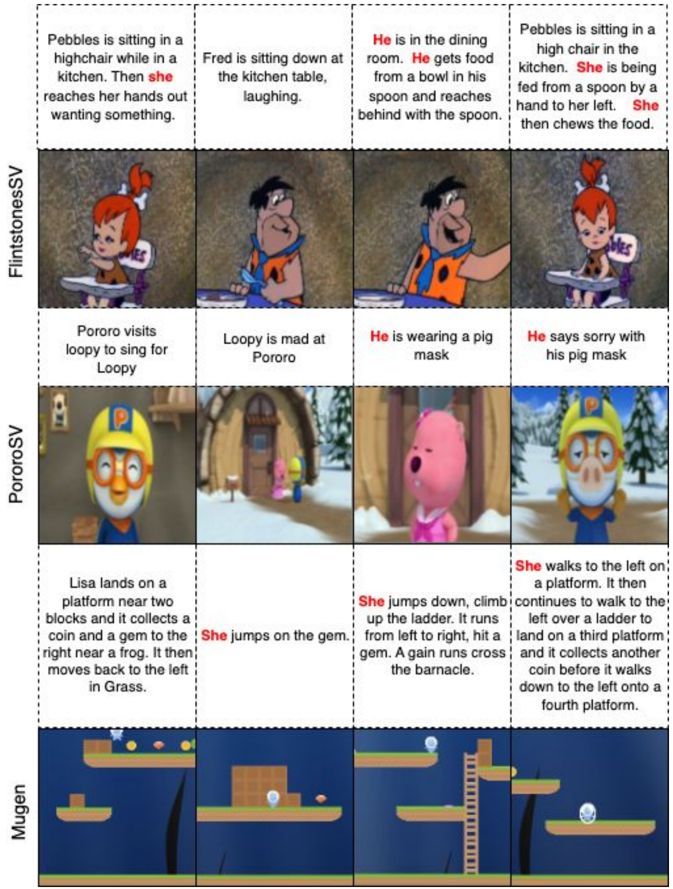
For each frame cross-attention: $C_{attn} = \sum_i \hat{f}(\mathbf{Z}^m)_i f(\mathbf{S}^m)_i$

Memory attention module: $M_{attn} = \sum_{k=1}^{m-1} \sum_i \hat{f}(\mathbf{Z}^k)_i f(\mathbf{S}^k)_i f(\mathbf{S}^m)_i$

Final Output of the attention-module: $C_{attn} + M_{attn}$

Dataset Statistics

Dataset	#Ref (avg.)	#Chars	#Backgrounds
MUGEN [1]	None	1	2
Extended MUGEN	3	3	6
FlintstonesSV [2]	3.58	7	323
Extended FlintstonesSV	4.61	7	323
PororoSV [3]	1.01	9	None
Extended PororoSV	1.16	9	None



Pebbles is sitting in a highchair while in a kitchen. Then she reaches her hands out wanting something.

Fred is sitting down at the kitchen table, laughing.

He is in the dining room. He gets food from a bowl in his spoon and reaches behind with the spoon.

Pebbles is sitting in a high chair in the kitchen. She is being fed from a spoon by a hand to her left. She then chews the food.

Pororo visits loopy to sing for Loopy

Loopy is mad at Pororo

He is wearing a pig mask

He says sorry with his pig mask

Lisa lands on a platform near two blocks and it collects a coin and a gem to the right near a frog. It then moves back to the left in Grass.

She jumps on the gem.

She jumps down, climb up the ladder. It runs from left to right, hit a gem. A gain runs cross the barnacle.

She walks to the left on a platform. It then continues to walk to the left over a ladder to land on a third platform and it collects another coin before it walks down to the left onto a fourth platform.

Evaluation Metrics

- Character Classification : Measure frame accuracy (*i.e.* exact matching of characters) and F1-score by using fine-tuned Inception-v3.
- Background Classification: Measure the correspondence of the background to the ground-truth and consider F1-score as a measure of quality.
- Frechet Inception Distance (FID): Measure the distance between feature vectors from real and generated images.

Quantitative Results

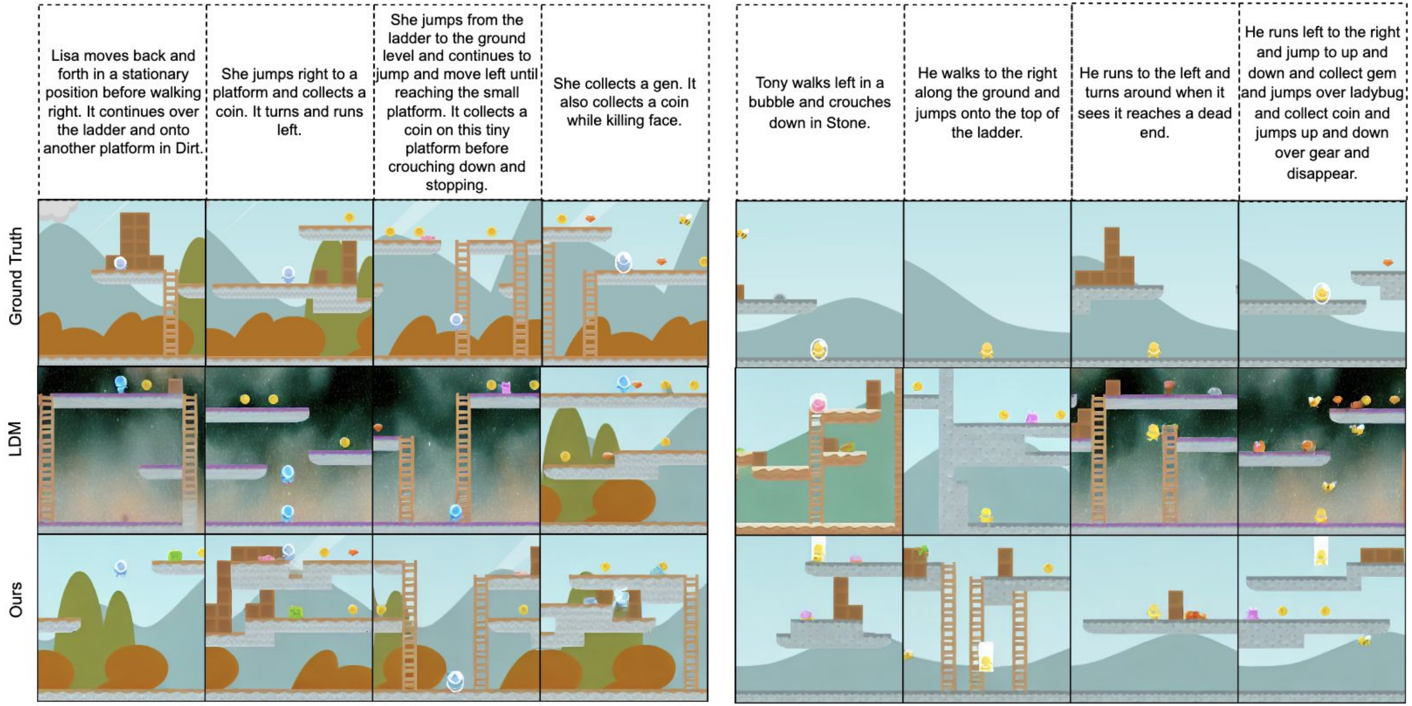
Dataset	Method	w/ref. text	Char-acc (↑)	Char-F1 (↑)	BG-acc (↑)	BG-F1 (↑)	FID (↓)
FlintstonesSV	VLCStoryGAN [4]	x	27.73	42.01	4.83	16.49	120.85
	LDM [5]	x	79.86	92.33	48.02	37.86	61.40
	LDM [5]	✓	57.38	78.68	44.19	28.25	87.39
	Story-LDM (Ours)	✓	69.19	86.59	35.21	28.80	69.49
PororoSV	VLCStoryGAN [4]	✓	17.36	43.02	-	-	84.96
	LDM [5]	✓	16.59	56.30	-	-	60.23
	Story-LDM (Ours)	✓	**20.26**	**57.95**	-	-	**36.64**
MUGEN	LDM [5]	✓	31.39	21.28	15.74	18.66	120.99
	Story-LDM (Ours)	✓	**93.40**	**95.60**	**92.19**	**92.37**	**62.16**

+22

+62

Improvement of **22%** and **62%** in terms of character accuracy on the PororoSV and the Mugen dataset.

Story Generation Results on MUGEN



Lisa moves back and forth in a stationary position before walking right. It continues over the ladder and onto another platform in Dirt.

She jumps right to a platform and collects a coin. It turns and runs left.

She jumps from the ladder to the ground level and continues to jump and move left until reaching the small platform. It collects a coin on this tiny platform before crouching down and stopping.

She collects a gen. It also collects a coin while killing face.

Tony walks left in a bubble and crouches down in Stone.

He walks to the right along the ground and jumps onto the top of the ladder.

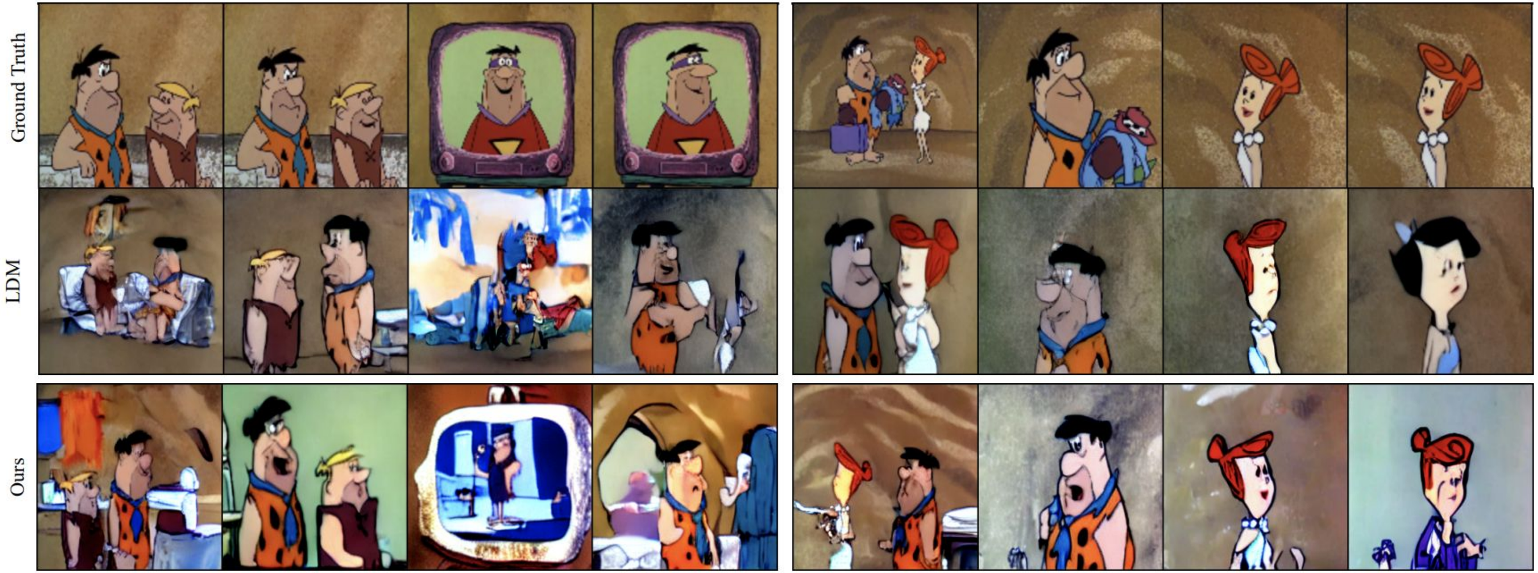
He runs to the left and turns around when it sees it reaches a dead end.

He runs left to the right and jump to up and down and collect gem and jumps over ladybug and collect coin and jumps up and down over gear and disappear.

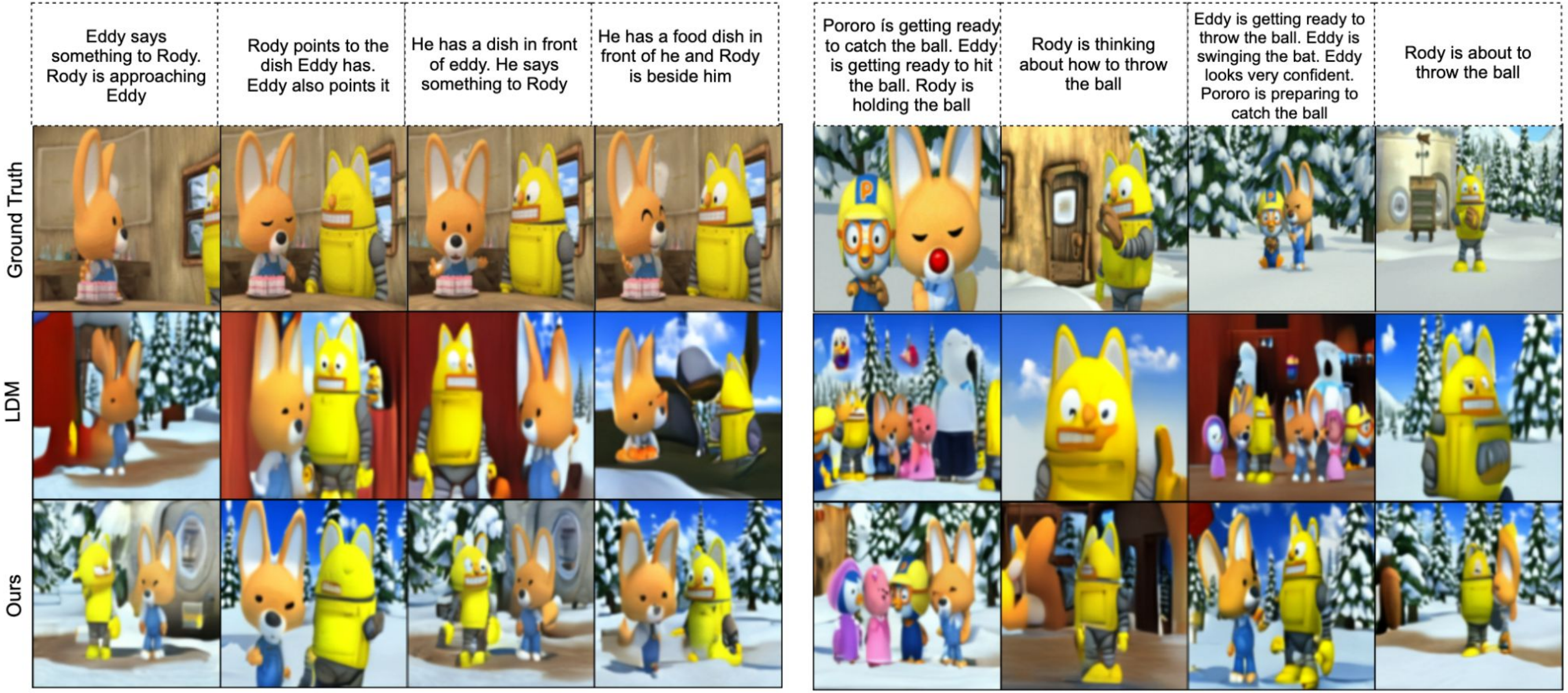
Story Generation Result on FlintstonesSV

Fred and Barney are having a conversation on the couch in the living room. Fred turns his head with a disdained look.
They are sitting in a room. Fred looks angry and sticks his tongue out while he and Barney are talking.
The circus performer on the television talks to his audience.
Fred is on tv in the room wearing a super hero outfit.

Wilma and Fred are standing in the room. Wilma is talking to Fred. Fred is holding a suitcase and clothing.
Fred is standing in a room speaking with an exasperated expression. He is holding a pile of clothes.
Wilma is standing in a room, she blinks twice before speaking.
She is standing in a room, talking to someone. Then, she giggles.



Story Generation Result on PororoSV



Eddy says something to Rody. Rody is approaching Eddy

Rody points to the dish Eddy has. Eddy also points it

He has a dish in front of eddy. He says something to Rody

He has a food dish in front of he and Rody is beside him

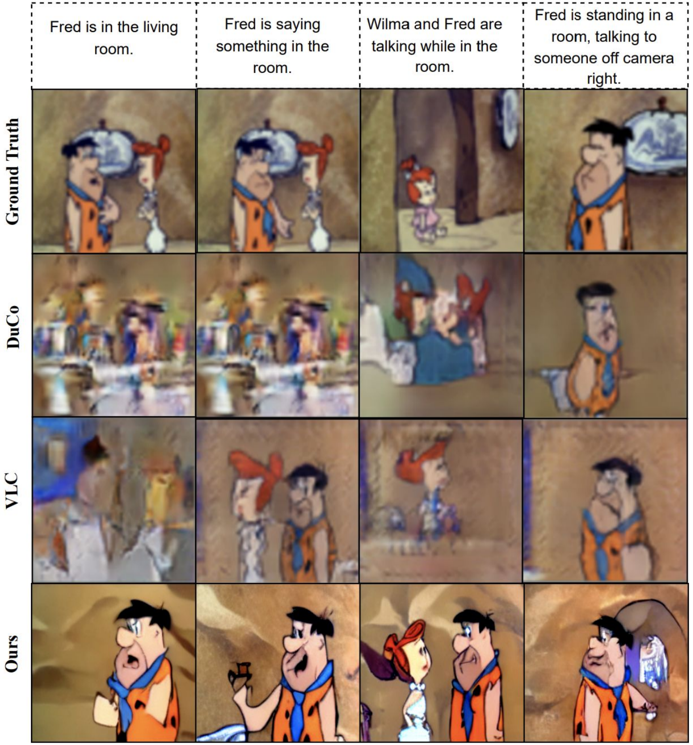
Pororo is getting ready to catch the ball. Eddy is getting ready to hit the ball. Rody is holding the ball

Rody is thinking about how to throw the ball

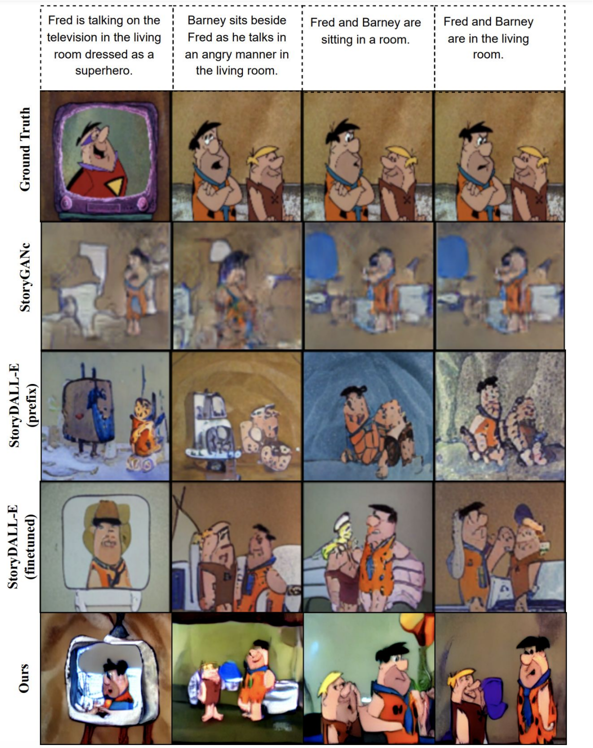
Eddy is getting ready to throw the ball. Eddy is swinging the bat. Eddy looks very confident. Pororo is preparing to catch the ball

Rody is about to throw the ball

Qualitative Comparison

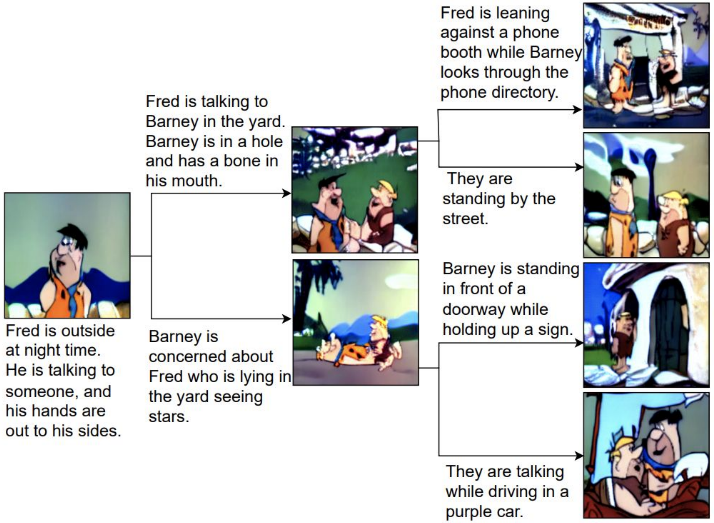

Fred is in the living room.
Fred is saying something in the room.
Wilma and Fred are talking while in the room.
Fred is standing in a room, talking to someone off camera right.
Ground Truth
DuCo
VLC
Ours

(a) Comparison on the FlintstonesSV dataset visual **story generation**.


Fred is talking on the television in the living room dressed as a superhero.
Barney sits beside Fred as he talks in an angry manner in the living room.
Fred and Barney are sitting in a room.
Fred and Barney are in the living room.
Ground Truth
StoryGANc
StoryDALL-E (prefix)
StoryDALL-E (finetuned)
Ours

(b) Comparison on the FlintstonesSV dataset for **story continuation.**

Additional qualitative results



Fred is outside at night time. He is talking to someone, and his hands are out to his sides.

Fred is talking to Barney in the yard. Barney is in a hole and has a bone in his mouth.

Fred is leaning against a phone booth while Barney looks through the phone directory.

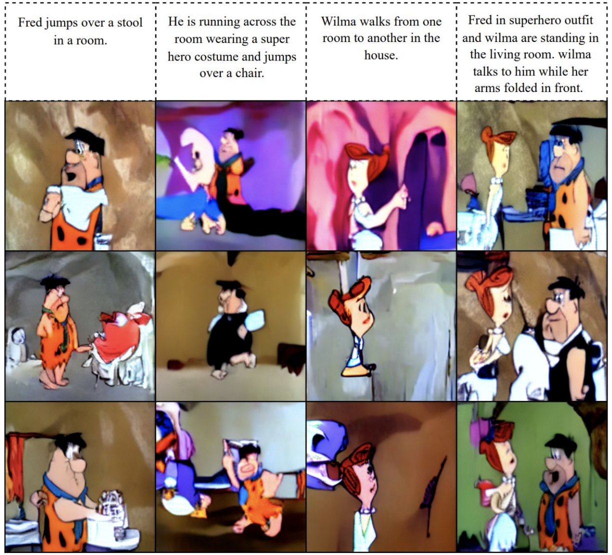
They are standing by the street.

Barney is concerned about Fred who is lying in the yard seeing stars.

Barney is standing in front of a doorway while holding up a sign.

They are talking while driving in a purple car.

(a) Generating different yet consistent stories by branching the storyline.



Fred jumps over a stool in a room.

He is running across the room wearing a super hero costume and jumps over a chair.

Wilma walks from one room to another in the house.

Fred in superhero outfit and wilma are standing in the living room. wilma talks to him while her arms folded in front.

(b) Diverse outputs for a single storyline obtained with our Story-LDM.

Conclusion

- We formulate consistent story generation in a more realistic way by co-referencing actors/backgrounds in the story descriptions.
- We introduce an autoregressive Story-LDM approach with memory attention capable of maintaining consistency across the frames.
- We expect our proposed formulation and models to be conductive to the real-world use cases and further the research.

References

[1] Hayes, Thomas, et al. "Mugen: A playground for video-audio-text multimodal understanding and generation." ECCV 2022.

[2] Gupta, Tanmay, et al. "Imagine this! scripts to compositions to videos." ECCV 2018.

[3] Li, Yitong, et al. "Storygan: A sequential conditional gan for story visualization." CVPR 2019.

[4] Maharana, Adyasha, and Mohit Bansal. "Integrating visuospatial, linguistic and commonsense structure into story visualization." arXiv preprint arXiv:2110.10834 (2021).

[5] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." CVPR 2022.