A. Details of Used Datasets

PororoSV PororoSV contains 10191/2334/2208 samples of the train, val, and test set, respectively. Each sample contains 5 consecutive frames sampled from videos. There are 9 main characters in PororoSV: Pororo, Loopy, Eddy, Harry, Poby, Tongtong, Crong, Rody, and Petty. Profile pictures of them are given in Fig. 8.

Figure 8. Main character names and corresponding photos in PororoSV. The photos are from https://pororo.fandom.com/

FlintstonesSV FlintstonesSV contains 20132/2071/2309 samples of the train, val, and test set, respectively. Each sample contains 5 consecutive frames sampled from videos. There are 7 main characters in PororoSV: Fred, Barney, Wilma, Betty, Pebbles, Dino, and Slate. Profile pictures of them are given in Fig. 9.

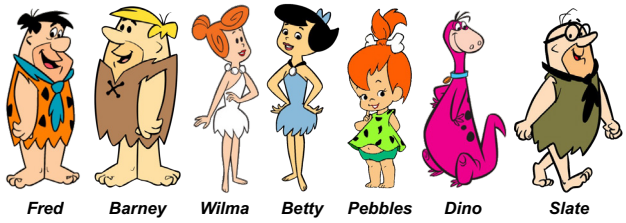


Figure 9. Main character names and corresponding photos in FlintstonesSV. The photos are from https://flintstones.fandom.com/

VIST VIST contains 23344/2921/2925 samples of the train, val, and test set, respectively. Each sample contains 5 consecutive frames and two kinds of captions: SIS and DII. It should be noted that VIST is sampled from image albums instead of video, so the visual stories are not as consistent as the ones in PororoSV and FlintstonesSV. However, the five images can still form a coherent story. There are no recurring characters in VIST. The VIST dataset used in this paper is a subset of the original one. Because some images are removed by their owners and are not accessible now, we drop the stories containing such images. We further choose stories that contain both SIS and DII captions.

B. Detailed Human Evaluation Settings

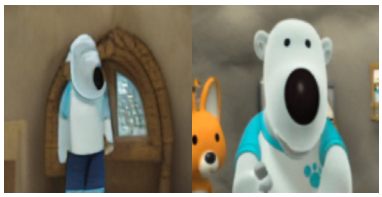
We provide human evaluation results regarding visual quality, relevance, and consistency. Human annotators tend to choose visual stories in high visual quality. This may confuse the three separate evaluation criteria. To make criteria orthogonal to each other, we carefully design the human evaluation process. Specifically, we only provide single images (without captions) for visual quality evaluation, single images and the corresponding specific captions for relevance evaluation, and whole visual stories (without captions) for consistency evaluation. For the detailed annotation standards and instructions given to annotators, see Appendix H.

C. Win and Lose Cases in Human Evaluation

In this section, we provide some cases in our human evaluation. Specifically, Fig. 10, Fig. 11, and Fig. 12 show the cases that AR-LDM wins StoryDALL·E regarding visual quality, relevance, and consistency, respectively. Fig. 13 show the cases that AR-LDM loses StoryDALL·E in human evaluation.

(a) Cases on PororoSV. (b) Cases on FlintstonesSV. (c) Cases on VIST-SIS.

Figure 10. Cases that AR-LDM **wins** StoryDALL·E in human evaluation regarding **visual quality**. The left ones are synthesized by StoryDALL·E, and the right ones are synthesized by AR-LDM.



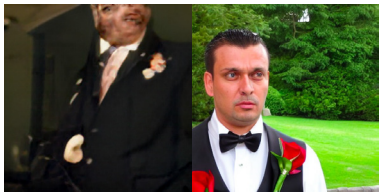
Poby stands up and give an advice.

Barney is in the car. He talks while Fred sits next to him.

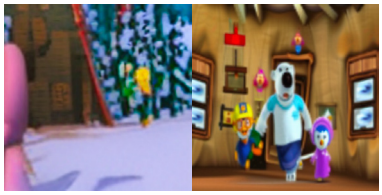
There were lots of people at the club.

On the bookshelf Petty take out a book about cooking.

Fred and Barney are standing on the sidewalk talking.



Here is the anxious groom waiting.

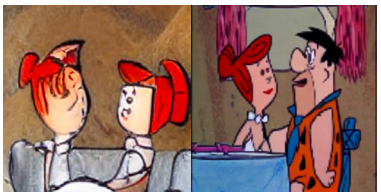


Poby and friends are running fast through the buildings.

Fred is in a room. He is talking and smiling.

Everyone did their best with the pumpkins.

King Harry told Tongtong to go out. Harry ordered seriously.



Fred and Wilma are sitting in a dining room. Fred speaks to Wilma.

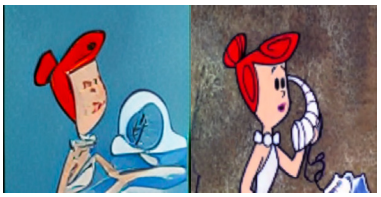
As well as Nate and Derrick two local DJs.

Pororo and Crong looks tired. It's snowing outside.

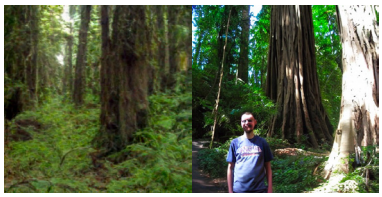
Fred wears a mask and holds a hammer while he stands in front of a sign outside.

The trail was really quite beautiful.

Pororo and Crong looks tired. It's snowing outside.



Wilma is in a room. She talks on the phone.



He's happy to be in the redwood forest.

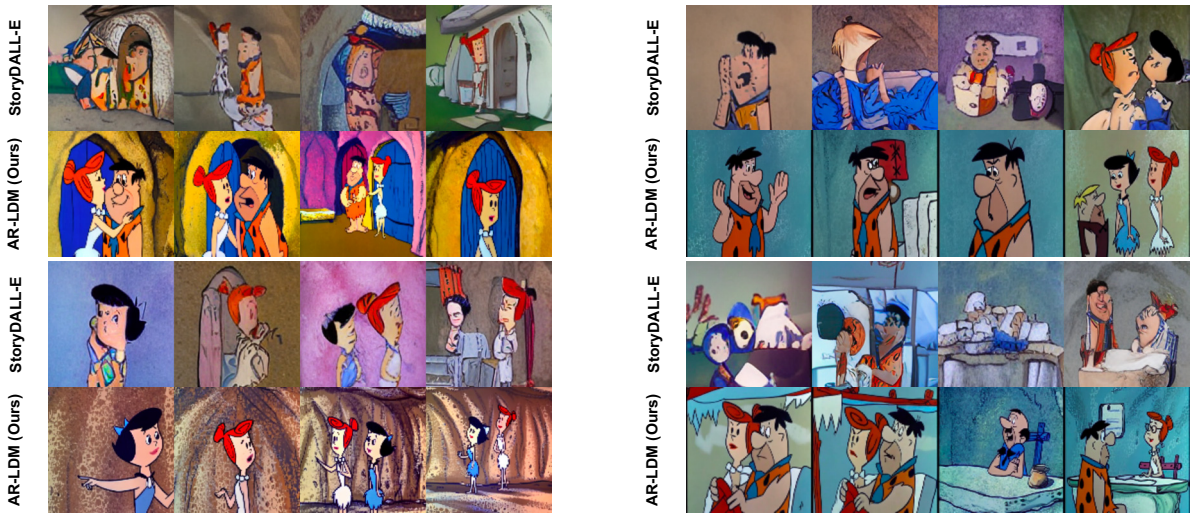
(a) Cases on PororoSV.

(b) Cases on FlintstonesSV.

(c) Cases on VIST-SIS.

Figure 11. Cases that AR-LDM **wins** StoryDALL·E in human evaluation regarding **relevance**. The left ones are synthesized by StoryDALL·E, and the right ones are synthesized by AR-LDM.

(a) Cases on PororoSV.



(b) Cases on FlintstonesSV.

(c) Cases on VIST-SIS.

Figure 12. Cases that AR-LDM **wins** StoryDALL·E in human evaluation regarding **consistency**.

(a) Cases regarding visual quality. The left ones are synthesized by StoryDALL·E, and the right ones are synthesized by AR-LDM.

(b) Cases regarding relevance. The left ones are synthesized by StoryDALL·E, and the right ones are synthesized by AR-LDM.

(c) Cases regarding consistency.

Figure 13. Cases that AR-LDM **loses** StoryDALL·E in human evaluation.

D. Additional Synthesized Visual Stories

In this section, we provide additional synthesized visual stories.

D.1. PororoSV

We provide additional cases on PororoSV in both story visualization (Fig. 14) and story continuation (Fig. 15) settings. The corresponding captions are listed below.

Case 1:

1. Petty Pororo and Poby arrives at Loopy's house.
2. Loopy opens door and invites Loopy friends in.
3. Pororo and Poby friends are finished with meal.
4. Poby gives the thumbs up.
5. Poby and Petty give the thumbs up. Loopy is happy.

Case 2:

1. Pororo and Poby friends are finished with meal.
2. Poby gives the thumbs up.
3. Poby and Petty give the thumbs up. Loopy is happy.
4. Loopy suggests to go outside.
5. Pororo agrees and smiles with Poby hand.

Case 3:

1. The weather is snowy and windy. There isn't any person in this scene.
2. The weather is snowy and windy and the weather is going even worse. Two characters are running toward the cabin.
3. Eddy is now at the room. Eddy uses pencil ruler and papers. Eddy seems satisfied with his work.
4. Eddy is in the room. Eddy uses pencil ruler and papers. Eddy turns his head right side.
5. Poby Petty Loopy and Harry are gathering in the cabin. The weather is snowy and windy. Poby thinks that emergency situation happens.

Case 4:

1. Eddy keep holding his picture and explains his idea to Poby. Because of the snowy weather we have no choice but to rescue Pororo and Crong by airship. Therefore Eddy resorts to Poby that they need Poby's help.
2. Poby seems surprised because Poby doesn't expect that Poby will be needed in this situation. After hearing from Eddy Poby turns his head to the left side.
3. Pororo and Crong are in the middle of the mountain. They seem tired and exhausted. Pororo close his eyes with long hard thinking.
4. Pororo closes his eyes with long hard thinking. The weather is snowy and it becomes worse. Pororo can't find any other solutions except rope to get out of this mountain. Pororo and Crong are stuck in this mountain so Pororo tries to use rope.
5. Pororo and Crong try to pull the rope to overcome this situation. However it is hard to fully apply their force.

Case 5:

1. In Loopy's imagination Pororo comes to her with flowers.
2. From far away Crong also comes to Loopy.
3. Loopy is wearing a fine costume and is holding a parasol. Crong gives her flowers.
4. Eddy with his mustache and with his car presents flowers to Loopy.
5. Loopy stands in front of the mirror and checks herself.

Case 6:

1. Poby feels ashamed and wants that nobody saw him falling down.
2. Seeing Poby through the telescope Eddy secretly smiles and talks to himself that Eddy saw Poby falling down.
3. Eddy is interested in seeing things and friends through telescope. Eddy brings telescope and goes to the mountain to observe his friends more.
4. Up on the mountain Eddy chooses a target. It is Pororo. Eddy looks through the telescope.
5. Pororo was reading a book. Loopy calls him outside his home. Pororo hears it and looks at the door if anyone came.

Case 7:

1. Pororo after looking Crong goes to bed says good night to Crong.
2. Pororo is lying on the bed.
3. The morning came. It got bright.
4. Crong tried but could not make number two.
5. Pororo called Crong from outside.

Case 8:

1. The car Pororo and Crong are on the ground.
2. There is Pororo's house in the forest.
3. Pororo's house in covered with snow.
4. Pororo and friends are eating lunch.
5. Eddy and Loopy are sitting next to each other.

Case 9:

1. The car tells Crong that Pororo and Crong are to meet at Eddy's house for a picnic.
2. Remembering the appointment Crong was surprised.
3. While Pororo is sleeping Crong calls Pororo.
4. Pororo and Crong came out of the house. Pororo and Crong ride in the car.
5. The car arrives at Eddy's house.

Case 10:

1. The car arrives at Eddy's house.
2. Pororo and Crong are coming out of the car.
3. Pororo and Crong greet friends friends.
4. All the friends are standing in front of Eddy's house.
5. Pororo is asking about something.

Case 11:

1. Poby and Harry face each other with smile. Harry looks excited.
2. Harry's house lays down on one side of the Poby's house.
3. Harry sits down on the bed. Harry really skips about for joy.
4. Harry sits down on her bed with joy.
5. Harry is looking out of the window.

Case 12:

1. Harry is looking out of the window.
2. Pororo and his friends sit round with joy. Cake lies on the table.
3. Harry stand up in front of Harry's house.
4. Pororo and his friends are sitting around the table. They congratulates Harry's new house.
5. Harry stands up beside the cake. Harry is really happy.

Figure 14. Example of generated visual stories (left 5 frames) from AR-LDM and corresponding ground truths (right 5 frames) on PororoSV. These cases are under **story visualization** setting.

Figure 15. Example of generated visual stories (left 5 frames) from AR-LDM and corresponding ground truths (right 5 frames) on PororoSV. These cases are under **story continuation** setting, which means the first frame serves as a source frame. These cases are corresponding to Fig. 14.

D.2. FlintstonesSV

As shown in Fig. 16, We provide additional cases on FlintstonesSV in the story continuation setting. The corresponding captions are listed below.

Case 1:

1. Fred and Wilma are standing in a room. Wilma speaks to Fred.

2. Fred and Wilma are in the room, Fred is talking. Wilma reaches to hug him.

3. Wilma is speaking in the room.

4. Wilma and Fred are in a room. Fred is grabbing Wilma by the hips and pushing her through the room.

5. Fred is standing in the dining room, waiting to be served.

Case 2:

1. Wilma and Betty are standing in a room speaking.

2. Wilma and Betty are standing in a room talking.

3. Barney turns to talk to someone behind him in the room with an angry look on his face.

4. A small boy holding the stone is in the living room. He holds the tv as he walks.

5. Barney is in the living room talking sternly and wagging his finger.

Case 3:

1. Fred and Barney are outside, standing next to a car. Fred holding money in his hand while speaking to someone.

2. Barney is outside pointing at something. While he is pointing he is saying something.

3. Fred is holding money in the room.

4. Fred looks at some money and talks in a store.

5. Betty and Wilma are sitting in a car. Wilma tugs at a rope while betty leans back in her seat.

Case 4:

1. Fred is outside. He talks with a tear in his eye.

2. Fred is in the backyard in front of the stone wall talking to someone.

3. Fred is standing outside while crying and talking.

4. Barney is in the yard. He is talking and gesturing with his hand.

5. Fred and Barney are sitting in a car talking.

Case 5:

1. Fred is awkwardly shaking a man's hand in the room.

2. Fred is driving a car down the road. He is speaking to Barney.

3. Wilma is holding a basket out in the yard. She is listening to Betty. They are standing in front of a stone fence.

4. Betty is standing in her yard, talking to someone off camera right.

5. Wilma is walking through the yard. She is carrying a basket. While walking she is speaking to someone behind her.

Case 6:

1. Wilma talking to Fred in a room.

2. Wilma talks to Fred with her hands on his back in the living room.

3. Wilma and Fred are in the room. Fred looks upset and says something. Wilma holds his shoulders and says something.

4. Fred and Wilma are standing in a room. Fred speaks while Wilma holds onto his shoulder.

5. A police officer in uniform with a long skinny nose is standing in a doorway talking to Wilma.

Case 7:

1. Barney laughs and talks to bamm bamm while they walk with Betty outside.

2. the animal is standing on a rock and clapping its fins in a cave.

3. Fred and Barney are in the car talking to each other.

4. Fred is driving a car while Barney rides in the passenger seat. They talk briefly.

5. The man in green is outside holding a bag and wearing a hat. He pushes the man with glasses in pink clothes and purple tie.

Case 8:

1. Betty is standing in a room. She speaks, leans back, and then begins to race off.

2. Wilma and Betty are standing on the driveway and looking inside the garage.

3. Wilma in the room talking to someone.

4. Barney walks through the yard in a pink shirt. He looks back while talking, then looks forward and laughs with his eyes closed.

5. The great gazoo floats in the room as he speaks.

Case 9:

1. Fred is in the living room. He is talking to someone.

2. Barney is sitting on a bench in a bowling alley while Fred stands and they have a conversation.

3. Barney is in a room talking.

4. first, Fred looks down at the bowling ball at the bowling alley with a funny confused look and sticking his tongue out. Then, Fred looks behind him while still holding the bowling bowl.

5. Fred is in the bowling alley. He stands with a while bowling ball in his hands and then runs to the left in preparation to bowl. He speaks to someone off screen to the right.

Case 10:

1. Fred is sitting on the ground in the dressing room. He is wearing a purple eye mask and red suit with a purple cape.

2. Barney is talking near the doorway.

3. Betty and Wilma are sitting in a living room. Wilma begins to cry and brings a tissue to her face. Then Betty turns to look at Wilma.

4. Wilma and Betty are sitting on a couch in the living room. Wilma speaks to Betty and cries into a handkerchief.

5. Wilma and Betty are sitting on a couch in the living room. Wilma is crying and wiping her tears with a handkerchief while Betty speaks to her.

Case 11:

1. The fancy man in white suit is in the room, pointing to the ceiling as he talks.

2. The musician with guitar is on stage. He is singing.

3. There is a man playing guitar on a stage.

4. The man with blue short playing guitar is on the stage. He is dancing.

5. Three men are on stage playing guitar and singing. There is a man with black hair, a man with brown hair, and a man with red hair.

Case 12:

1. Wilma is sitting in the living room and talking to Fred.

2. Fred and Wilma are in the living room. Wilma is angry and Fred is talking to her.

3. Fred, Wilma, and Pebbles are in the living room. Fred stands in front of Wilma, who is standing on the couch. Fred speaks to Wilma while Pebbles sits on the floor, playing with a stick.

4. Fred is kneeling on the floor in a room while watching Pebbles play.

5. Pebbles is on her hands and knees on the floor of a room. She blinks her eyes and then lowers her head.

Figure 16. Example of generated visual stories (left 5 frames) from AR-LDM and corresponding ground truths (right 5 frames) on FlintstonesSV. These cases are under **story continuation** setting, which means the first frame serves as a source frame.

D.3. VIST-SIS

As shown in Fig. 17, We provide additional cases on VIST-SIS in the story continuation setting. The corresponding captions are listed below.

Case 1:

1. I went to the wedding last week.
2. It was on the lake.
3. I brought all the necessary paperwork.
4. The live band was very good.
5. I bought many flowers for the couple.

Case 2:

1. On the night of the party everyone was so excited to see each other.
2. A few of the guys broke out the guitar and started to play some tunes.
3. My friend james got a photo standing next to his favorite character darth vader.
4. For most of the night we decided to play retro video games.
5. The gun I used while playing the nintendo game dunk hunt.

Case 3:

1. Fixing up his bike.
2. Following the pacific coast bike route.
3. Taking a lunch break.
4. Eating fruits and sandwiches.
5. About to cross the san francisco bridge.

Case 4:

1. From the entrance, the decorations told us we were having a traditional japanese meal.
2. The low tables had all sorts of delicious food our hosts had prepared.
3. But that was just the beginning, we saw as we walked towards a larger table with a fest spread upon it.
4. Soon we'd consumed everything – it was all so tasty, and interesting, too.
5. We took a moment before we left to take a closer look at the art on the walls, and found the meaning for our hosts; I guess it's rude to ask to come back right away!

Case 5:

1. The three sisters gathered to celebrate thanksgiving.
2. Their mom made a lot of good food for dinner.
3. She even made tons of mashed potatoes and spaghetti sauce.
4. The plate of pulled pork is always a crowd pleaser.
5. Black eyed peas are delicious as well.

Case 6:

1. The city is very crowded with people.
2. The cops watch over to keep it safe.
3. We are getting ready to set up the party.
4. We are carrying the table.
5. Ken is adjusting everything just right.

Case 7:

1. A bright blue sky to start off the day.
2. Taking the trike on a drive across town.
3. Passed an odd shaped trash can that could have seen better days.
4. Saw a beautiful glass sign that was perfect for the old lady but the owner refused to sell.
5. Above a crow watches from the chimney.

Case 8:

1. When I went out for a walk this morning I was taken aback by just how beautiful the snow was.
2. We get snow back home, but nothing like how this snow just seems to go on forever.
3. I was happy to see others out enjoying the snow.
4. The roadway was so nicely cleared.
5. And you couldn't beat watching the sunrise in such a place.

Case 9:

1. We visited some historic sites on our trip.
2. They had a lot of warning signs when we arrived.
3. After that we traveled the paths at the site.
4. Then we came across signs that told us the history of the site.
5. After that we came across some old graves at the site.

Case 10:

1. Before going to my storage locker I stopped off for some food.
2. This place always serves up great sandwiches.
3. After eating I made it the storage unit.
4. It is so full of papers and other things that I do not need.
5. I found this old desk in there that I forgot I had.

Case 11:

1. I love going to the art fest.
2. Me and my mom got our faces painted.
3. My daughter was transformed with the face paint.
4. So many people had booths there.
5. My husband even had a good time.

Case 12:

1. Greetings from the disneyland Halloween party!
2. I dressed like a goth.
3. Surprisingly, I got a fair amount of candy!
4. I also saw so spooky ghost dancers,
5. And even count mickey! hope to see you thanksgiving!

Figure 17. Example of generated visual stories (left 5 frames) from AR-LDM and corresponding ground truths (right 5 frames) on VIST-SIS. These cases are under **story continuation** setting, which means the first frame serves as a source frame.

D.4. VIST-DII

As shown in Fig. 18, We provide additional cases on VIST-DII in the story continuation setting. These cases are corresponding to Fig. 17, the captions are listed below.

Case 1:

1. A man walking on a red carpet in an outdoor building set up for a wedding.
2. A man and a woman pose together, she in a black dress, he in a red tie.
3. The lady that performed the wedding ceremony was a native american.
4. An entertainer performing music and singing at an event.
5. A bouquet of flowers features white flowers and succulent greens.

Case 2:

1. Group photo with a group of females wearing white tops and sunglasses.
2. One man is playing the guitar while the other watches on.
3. Very happy young man getting his picture taken with darth vader.
4. Long haired man lying on the floor playing a videogame.
5. A nintendo gun-shaped video game controller lays on a burnt orange carpet.

Case 3:

1. A young man is working on fixing his pedal bike.
2. A sign states that a bike route to the pacific coast of california is up ahead.
3. To people wearing helmets standing with their bicycles.
4. A person wearing an orange jacket and helmet stands next to a bicycle and picnic table.
5. A woman bicyclist posing in front of the san francisco bay bridge.

Case 4:

1. This space has three pictures hanging on the wall and a plant in a vase and two small sculptures on a table.
2. Square table with food on top in bowls and containers.
3. A very big meal all laid out on a long table.
4. A tray holding dishes and a tea kettle sits on a table set with food.
5. A wall has scrolls hanging down it and flowers are next to it.

Case 5:

1. Three woman holding wine in the foyer of a nice house.
2. Plates of food sitting on a kitchen table in front of window with white blinds.
3. Two bowls that are filled with different types of foods.
4. A plate of piled cooked crispy bacon ready to be eaten.
5. A crock pot filled with delicious looking warm soup.

Case 6:

1. A mob of people wait for the light to change to walk across a city street.
2. Small silver building on a busy street with one guard and a us army advertisement.
3. Workmen work at something below a marine advertisement.
4. Construction workers making repairs on the side of the road in a big city.
5. A worker repairing a window frame of the u.s. armed forces career center.

Case 7:

1. The bars of the structure are yellow and the sky is extremely clear.
2. A man in helmet rides a motorcycle with three wheels.
3. A black and white image of a metal grated trash can sitting on a brick sidewalk.
4. A section of a stained glass window displaying a blue oval with a white stripe detailed with plant like shapes.
5. An old building with a brick chimney in marginal shape, with a crow perched on top.

Case 8:

1. A road filled with snow with two trees that have no leaves.
2. The white snow is breathtaking to coexist with the sky.
3. A man pushes someone on a sled in this wintry scene.
4. Snow is covering the ground and the leafless trees.
5. The bare branches of a large tree are silhouetted against the sunset.

Case 9:

1. A green sign with white lettering describing the hours and rules for the D.A.R. historic site.
2. The sign was displayed on a poster so all could see.
3. A dirt road runs through the forest of green trees.
4. The fort wilkinson sign located in the state of georgia.
5. A carving in stone gives remembrance to the citizens during the american revolution.

Case 10:

1. The outside of a panera bread restuarant is lit up at night.
2. A club sandwich sits beside a bowl of soup in this tempting meal.
3. A pile of items, including cartoon character cutouts and signs, is placed in a row of storage lockers.
4. A storage unit filled numerous boxes filled with various items.
5. A wooden desk sitting in the hall way of self storage.

Case 11:

1. A woman with blonde hair is giving another woman face paint at a fairground.
2. One woman leans into the ear of another while they are wearing face paint.
3. A girl with a dark complexion is getting her face painted for a special event.
4. A man is leaning to the side while sitting in a folding chair.
5. A man and a woman dressed as a clown pose in front of artworks displayed beneath canopies.

Case 12:

1. People walk through a festival in the street while a sign advertises a mickey mouse Halloween party.
2. The person is dressed up like a zombie on the walking dead.
3. Random fun size candy, skittles, tootsie roll, milky way, 3 musketeers, lemon heads, m&m's.
4. People in costumes are dancing on a ballroom floor.
5. A person in a mickey mouse costume stands by a red curtain.

Figure 18. Example of generated visual stories (left 5 frames) from AR-LDM and corresponding ground truths (right 5 frames) on VIST-DII. These cases are under **story continuation** setting, which means the first frame serves as a source frame. These cases are corresponding to Fig. 17

E. Additional Unseen Character Adaptation Results

In this section, we provide additional cases for adaptive AR-LDM. As shown in Fig. 20 and Fig. 21, AR-LDM can successfully adapt to the new character given only 3-5 images. In Fig. 19, We also present the training images and captions we used in the adaptation cases in Fig. 7, Fig. 20, and Fig. 21, respectively.

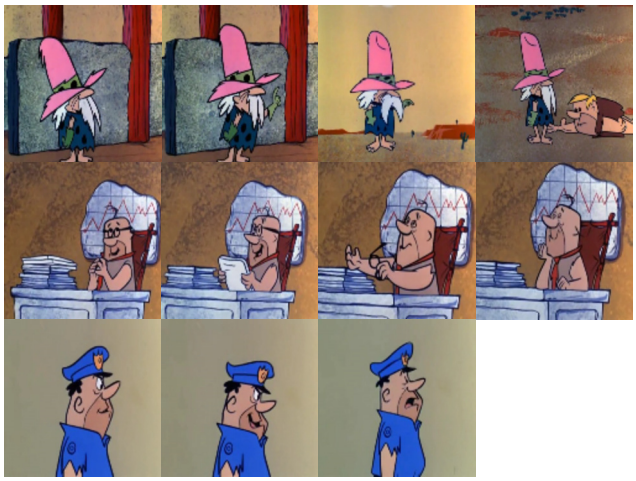


Figure 19. The training images used in the adaptation cases in Fig. 7, Fig. 20, and Fig. 21.

Case in Fig. 7:

1. <char> stands in the room speaking.

2. <char> is in the room. He talks as he points with his finger.

3. <char> is outside.

4. <char> is outside talking down to Barney who s laying on the ground with one of his hands stretched out above his head.

Case in Fig. 20:

1. <char> is sitting at his desk in the office and talking.

2. <char> speaks to himself in a office as he reads a piece of paper.

3. <char> is looking at the papers in his office speaking.

4. <char> ponders something at his office desk.

Case in Fig. 21:

1. <char> is outside talking.

2. <char> is outside. He speaks quickly.

3. <char> is talking outside.

Figure 20. Adaptation results for a case AR-LDM failed to properly generate on FlintstonesSV. The underlined texts refer to one specific person and can be replaced by <char> in adapted AR-LDM.

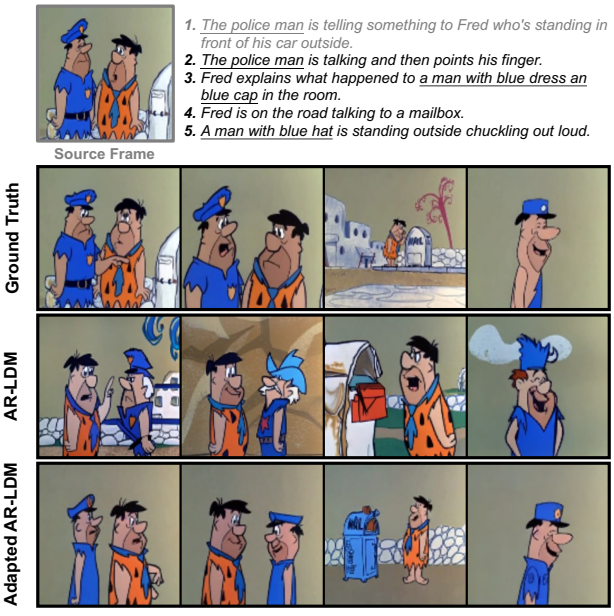


Figure 21. Adaptation results for a case AR-LDM failed to properly generate on FlintstonesSV. The underlined texts refer to one specific person and can be replaced by <char> in adapted AR-LDM.

F. Discussion

We find that many failure cases can be attributed to failing to ground the entity correctly. As shown in Fig. 22 (the captions are also given below), though AR-LDM is able to preserve the consistency of the bed across the second and the third frames, it also draws some entities in the wrong color. In the third frame, the blanket incorrectly grounds the green dress Wilma wears in the second frame. These failure groundings cause inconsistency across frames. We assume this can be attributed to BLIP's lack of the masked image modeling pre-training task. We suggest using multimodal PTMs with this pre-training task like BEiT-3 [36] after their weight is available.

(a) Ground truth visual story.

(b) Visual story synthesized by finetuned Stable Diffusion.

(c) Visual story synthesized by AR-LDM.

Figure 22. A failure case of AR-LDM.

1. Mr Slate who is really Fred is in his bedroom talking with his hand on his chest and then he points his finger.

2. Wilma is in the bedroom. She is sitting in the bed.

3. Fred and Wilma are in the bedroom. Wilma is beneath a blanket while Fred is beside the bed, speaking to her.

4. Fred is in the living room. He is talking and wearing a green shirt.

G. Ethical Statement

For the two datasets introduced in previous works, PororoSV and FlintstonesSV, images are unable to be confused with real images due to the fact that they are cartoons. As for the newly adopted VIST dataset, we follow a CreativeML Open RAIL-M license (https://huggingface.co/spaces/CompVis/stable-diffusion-license) used by Stable Diffusion [26]. Also, based on the pre-trained Stable Diffusion weight, we share the same biases and content acknowledgment as follows:

Biases and Content Acknowledgment Despite how impressive being able to turn text into image is, beware to the fact that this model may output content that reinforces or exacerbates societal biases, as well as realistic faces, pornography and violence. The model was trained on the LAION-5B dataset, which scraped non-curated image-text-pairs from the internet (the exception being the removal of illegal content) and is meant for research purposes.

H. Annotation Instructions for Human Evaluation

The annotation instructions for the human evaluation are provided here:

We have collected synthesized visual stories from different generative models. We would like you to help us choose the preferred ones from these stories regarding three orthogonal criteria, visual quality, relevance, and consistency. Visual stories from (anonymous) generative models have been shuffled on EACH ROW, so you the annotator cannot know which model they come from.

PLEASE READ THESE INSTRUCTIONS IN FULL.

Annotation Rules for Visual Quality:

- You are given two single images (without caption) synthesized by two different generative models to describe one specific caption. Choose the synthesized image only according to which one is in higher visual quality.
- What is the best answer? Make a decision based on (a) the scenes, entities, and characters are logical and not obviously unreasonable, like misplacement or clipping; (b) the image is clear, rich in detail, and realistic in color.
- If two images provide the same visual quality by your judgment, and there is no clear winner, you may rank them the tie, but please only use this sparingly.

Annotation Rules for Relevance:

- You are given one specific caption and two images synthesized by two different generative models to describe the given caption. Choose the synthesized image only according to which one is more relevant to the given caption.
- What is the best answer? Make a decision based on (a) the image conforms well to the content of the caption; (b) the image reflects the details and objects described in the caption instead of omitting them. Overall, use your best judgment to choose answers based on being the most relevant image, which we define as one which is at least somewhat correct, and minimally informative about what the caption is describing.
- Images in higher quality are not always the best. An image that is more relevant to the given caption may be better than one in higher quality, if they are at least as correct and informative.
- If two images provide the same relevance by your judgment, and there is no clear winner, you may rank them the tie, but please only use this sparingly.

Annotation Rules for Consistency:

- You are given two visual stories synthesized by two different generative models to describe specific captions. Choose the synthesized visual story only according to which one is more consistent.
- What is the best answer? Make a decision based on the scenes as well as recurring entities and characters are consistent across frames.
- Visual stories in higher quality are not always the best. A visual story that is more consistent across frames may be better than one in higher quality, if they are at least as correct and informative.
- If two visual stories provide the same consistency by your judgment, and there is no clear winner, you may rank them the tie, but please only use this sparingly.