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

Tanzila Rahman\(^1,3\) Hsin-Ying Lee\(^2\) Jian Ren\(^2\) Sergey Tulyakov\(^2\)
Shweta Mahajan\(^1,3\) Leonid Sigal\(^1,3,4\)
\(^1\)University of British Columbia \(^2\)Snap Inc.
\(^3\)Vector Institute for AI \(^4\)Canada CIFAR AI Chair

| Method               | Reference-text | Char-acc (↑) | Char-F1 (↑) | BG-acc (↑) | BG-F1 (↑) | FID (↓) |
|----------------------|----------------|--------------|-------------|------------|-----------|---------|
| Story-LDM (Ours)     | ✓              | 69.19        | 86.59       | 35.21      | 28.80     | 69.49   |
| Story-LDM (Ours)     | ×              | 83.29        | 94.61       | 35.54      | 27.32     | 64.89   |

Table 3. Experimental results using our proposed Story-LDM with and without the reference text on the FlintstonesSV dataset.

1. Additional Quantitative Results

In Tab. 3, we show the performance of our Story-LDM approach with and without the reference text. We observe that without reference text i.e., with explicit mentions of the characters (by their name), our approach outperforms the prior state-of-the-art VLCStoryGAN [32] (cf. Tab. 2, row 1 in the main paper) and the strong LDM [42] baseline on character accuracy by nearly ∼ 3.4%.

This demonstrates that introduction of our Memory attention module to the pipeline of the diffusion model with U-Net architecture increases the performance even in the traditional dataset setting. This can be attributed to the fact that even when textual resolution of character names or setting is unnecessary, our memory attention module can still enhance consistency of appearance in the visual domain. When using the descriptions with references, the character accuracy of the model drops by ∼ 12% showing the difficulty of the extended task of generating stories from the co-referenced text. Even the strong LDM baseline which outperforms the prior state-of-the-art VLCStoryGAN [32] (cf. Tab. 2 in the main paper) and the strong LDM [42] baseline on character accuracy by nearly ∼ 3.4%.

To evaluate fore- and back-ground consistency we propose two evaluation metrics: character and background classification. The generated story is considered accurate (positive sample) if it appropriately resolves character and background references in generation, otherwise it is considered a negative sample. In the figure, the 1st row (i.e. positive sample) has 100% character and background accuracy; the 2nd row (i.e. negative) has 75% and 50% character and background accuracy.

Fig. 10 shows examples of positive/negative samples. To evaluate fore- and back-ground consistency we propose two evaluation metrics: character and background classification. The generated story is considered accurate (positive sample) if it appropriately resolves character and background references in generation, otherwise it is considered a negative sample. In the figure, the 1st row (i.e. positive sample) has 100% character and background accuracy; the 2nd row (i.e. negative) has 75% and 50% character and background accuracy.

Fig. 10 shows examples of positive/negative samples. To evaluate fore- and back-ground consistency we propose two evaluation metrics: character and background classification. The generated story is considered accurate (positive sample) if it appropriately resolves character and background references in generation, otherwise it is considered a negative sample. In the figure, the 1st row (i.e. positive sample) has 100% character and background accuracy; the 2nd row (i.e. negative) has 75% and 50% character and background accuracy.

2. Additional Qualitative Results

In Fig. 9 we provide example image frames from our extended MUGEN dataset with three characters Lisa, Tony and Jhon, and six different backgrounds. Our extended MUGEN dataset is thus more complex than the MUGEN dataset in [16] where only one character is considered for two backgrounds.

Fig. 10 shows examples of positive/negative samples. To evaluate fore- and back-ground consistency we propose two evaluation metrics: character and background classification. The generated story is considered accurate (positive sample) if it appropriately resolves character and background references in generation, otherwise it is considered a negative sample. In the figure, the 1st row (i.e. positive sample) has 100% character and background accuracy; the 2nd row (i.e. negative) has 75% and 50% character and background accuracy.

Figs. 11, 12 and 13 show random samples obtained for story generation with our Story-LDM approach on the MUGEN, FlintstonesSV and PororoSV datasets respectively. Clearly, our approach yields high quality frames with character and background consistency.
Table 4. **Quantitative results.** Experimental results on the PororoSV datasets.

| Dataset       | Method                             | w/ ref. text | Char-acc (↑) | Char-F1 (↑) | BG-acc (↑) | BG-F1 (↑) | FID (↓) |
|---------------|------------------------------------|--------------|--------------|-------------|------------|-----------|---------|
| PororoSV      | DUCO-STORYGAN [32]                 | ✓            | 13.97        | 38.01       | -          | -         | 96.51   |
|               | VLCStoryGAN [31]                   | ✓            | 17.36        | 43.02       | -          | -         | 84.96   |
|               | LDM [42]                           | ✓            | 16.59        | 56.30       | -          | -         | 60.23   |
|               | Story-LDM (Ours)                   | ✓            | **20.26**    | **57.95**   | -          | -         | **36.64**|

Table 5. **Experimental results using non-referential text on the Flintstones and Mugen dataset.**

| Dataset       | Method                             | w/ ref. text | Char-acc (↑) | Char-F1 (↑) | BG-acc (↑) | BG-F1 (↑) | FID (↓) |
|---------------|------------------------------------|--------------|--------------|-------------|------------|-----------|---------|
| Flintstones    | VLCStoryGAN [31]                   | ×            | 27.73        | 42.01       | 4.83       | 16.49     | 120.85  |
|               | LDM [42]                           | ×            | 79.86        | 92.33       | 48.02      | 37.86     | 61.40   |
|               | Story-LDM (Ours)                   | ×            | **83.29**    | **94.61**   | 35.54      | 27.32     | 64.89   |
| Mugen         | LDM [42]                           | ×            | 95.25        | 97.04       | 21.10      | 23.98     | 123.69  |
|               | Story-LDM (Ours)                   | ×            | **97.60**    | **98.44**   | **74.72**  | **80.51** | **79.41**|
Figure 10. Positive (1st row) and negative (2nd row) examples.
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 jumps to up and down and collect gem and jumps over ladybug and collect coin and jumps up and down over gear and disappear.

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 the tiny platform before crouching down and stopping.

Jhon runs from right to left in Snow. He jumps down right on a gem and runs right from left to jumps upon a worm. Then it runs left from right to jump up to jumps down. He runs right to collect a coin before jumping upward but finally it moves without jumping. He runs from left to right. It jumps to on a face. He jumped down from a platform, looked confused running back and forth.

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 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 and land on a third platform and it collects another coin before it walks down to the left onto a fourth platform.

Jhon runs from left to right and jumps down a ledge in Planet. He paces back and forth. He avoids the bee by moving around. He lands on the ground and does two jumps to get onto a platform. It does a right jump and lands back down and runs onto some boxes and right jump to kill a face and collected a coin. It then runs right and collects another coin. He walks right to left across three boats, eating a coin on the first, a gem on the second and a coin on the third.

Figure 11. Qualitative results on the MUGEN dataset [16].
Figure 12. Qualitative results on the FlintstonesSV dataset [14].
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. Petty also said good bye to Poby and Harry. Petty was holding a piece of pie. Petty tried her pie. While Poby and Harry going back to their house they heard a voice calling them behind. Pororo and friends came to Poby and Harry. They were about to play hide and seek. Petty was holding a piece of pie. Petty tried her pie. While Poby and Harry going back to their house they heard a voice calling them behind. Pororo and friends came to Poby and Harry. They were about to play hide and seek.

**Figure 13.** Qualitative results on the PororoSV dataset [27].