

A. Experimental Setup

A.1. Datasets

We conduct our experiments on two dataset, StorySalon [13] and FlintStonesSV [17]. The StorySalon dataset [13] is the largest and most recent benchmark for visual storytelling. It contains 159,778 animation-style images across 446 character categories. Designed to support long-range story generation, each story consists of an average of 14 frames, and each corresponding text prompt contains an average of 106 words. The dataset is constructed from E-books and YouTube videos retrieved using keyword-based queries. To generate the narrative-aligned text descriptions for each image, TextBind [12] is applied to produce captions conditioned on both the image and the surrounding narrative text. Due to a portion of the YouTube videos becoming unavailable, we conduct all experiments on the E-books partition, which remains complete and high-quality. It includes 8,635 training stories with 118,892 frames, and 451 test stories with 6,026 frames.

The FlintstonesSV dataset [17] serves as an additional benchmarking dataset in our experiments. It comprises 25,184 densely annotated, animation-style video clips sourced from the animated series “The Flintstones.” FlintstonesSV selects a single representative frame from each clip and groups frames from consecutive clips into coherent stories of length five. The dataset consists of seven recurring characters, offering a robust setting for evaluating character and scene consistency in visual storytelling tasks. FlintstonesSV is split into 20,132 training, 2,071 validation, and 2,309 test stories.

A.2. Evaluation Metrics

To evaluate text-image alignment, we utilize CLIP text-image similarity (CLIP-T [21]) and TIFA [9] score. Following the literature [4, 13, 19], we use CLIP-T to evaluate high-level semantic text-image similarity. We also propose to use TIFA to assess the how well the generated images align with the narrative story text. Following [9], we use GPT-3.5 [2] to generate question-answer pairs and use UnifiedQA [11] to evaluate generated images.¹

To evaluate the quality of the generated images with respect to the ground-truth, we use Frechet Inception Distance (FID [7]) score and CLIP image-image similarity (CLIP-I [21]). FID is used to evaluate the quality of generated images by comparing the distributions of generated images and ground-truth images, quantifying the faithfulness and diverseness of generated images. CLIP-I measures the similarity between the generated and ground-truth images.

¹<https://github.com/Yushi-Hu/tifa>

A.3. Baselines

For StorySalon dataset, we compare our method with the following: (1) SDXL-Prompt:² We use the pretrained Stable Diffusion XL model [20] given the current text as prompt with a “A cartoon style image” prefix, without using any history image or text. (2) IP-Adapter [36]: We use the IP-Adapter for SDXL³ to generate the image given the current text as its prompt, and all the previous generated history images as its reference images. (3) State-of-the-art StoryGen [13]: We run the inference using the public checkpoint of the StoryGen model.⁴ For FlintStonesSV dataset, we compare our method with SDXL-Prompt, IP-Adapter and StoryGPT-V [27].

While a wide range of methods exist for visual storytelling, many rely on assumptions or constraints that are incompatible with our broader setting.

Specifically, we compare against StoryGen on the StorySalon dataset and StoryGPT-V on the FlintstonesSV dataset, which are, to the best of our knowledge, the only methods that operate under our broader setting.

Other methods were not included for the following reasons: (1) StoryGPT-V leverages predefined recurring characters specific to the FlintstonesSV dataset, a feature not available in the StorySalon dataset. This makes it inapplicable for evaluation on StorySalon. (2) We also excluded certain training-free methods (e.g., StoryDiffusion) because they are designed for short, “character + activity” format prompts and rely on repeated character prompt for consistency. We did experiments on StorySalon dataset with StoryDiffusion and OnePromptOneStory methods in our setting but yielded poor results. We believe that such comparisons would not be fair or meaningful.

Therefore, to our best knowledge, we chose the baselines that do not make such assumptions and are therefore applicable to the datasets we compare against.

A.4. Implementation Details

Our multi-modal history fusion model consists of $d = 4$ blocks, each block contains a cross-attention layer and a FeedForward Network. The hidden dimension of the fusion model is 1024. We use OpenCLIP-ViT-H [3] as the image and text encoders. Our history adapter is built on the Stable Diffusion XL. We train the model end-to-end for 80,000 steps with a batch size of 4. We use AdamW optimizer [15] with a learning rate of $1e-4$. During inference, we adopt a DDIM sampler [29] with 50 inference steps, and the guidance scale is 5. All experiments are conducted on

²<https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0>

³https://huggingface.co/h94/IP-Adapter/tree/main/sdxl_models

⁴https://huggingface.co/haoningwu/StoryGen/tree/main/checkpoint_StorySalon



Figure 5. ViSTA sample on FlintStonesSV test set.

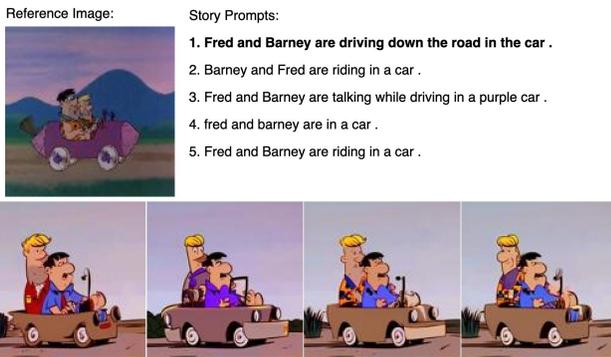


Figure 6. ViSTA sample on FlintStonesSV test set.

4 NVIDIA L40S GPUs. We set $\lambda = 0.5$ to balance image quality and consistency. The larger λ improves consistency but reduces quality, while the smaller λ improves quality at the cost of consistency. The value 0.5 offers the best trade-off in our experiments.

B. Additional Results

B.1. StorySalon Dataset

Additional qualitative results from the StorySalon dataset are presented in Figures 9, 10, 11, and 12. Consistent patterns are observed across different stories, including those featuring both human characters and animals, demonstrating that ViSTA achieves superior character and style consistency compared to other state-of-the-art baseline methods.

B.2. FlintStonesSV Dataset

Figures 5 and 6 illustrate examples generated by ViSTA from the FlintstonesSV dataset. The results demonstrate ViSTA’s effectiveness in maintaining consistent character identity and coherent visual storytelling.

C. Human Evaluation

To evaluate the quality of the generated story sequences, we conducted a pilot human study with five graduate student participants. Each participant was asked to assess the outputs from four different story generation methods.

The evaluation began with a set of detailed instructions provided through the survey interface, as shown below. These instructions outlined the evaluation procedure and defined the three core criteria: textual alignment, image alignment, and consistency.

Story Generation Methods Assessment

In this survey you will compare 4 different story generation methods that use state-of-the-art stable diffusion models across 5 examples. All methods start with a single prompt with its corresponding image as reference. After that, they generate a single image for each subsequent story prompt. For each example you will see the reference prompt-image pair and the subsequent story prompts, in addition to the original story book for reference on how the story looks like in a real-life example. You will be asked to compare the generated story based on 3 factors/metrics:

1. Textual Alignment: how well does each frame follow its corresponding prompt
2. Image Alignment: how well does each frame follow the overall style of the reference image
3. Consistency: how well is the frame-to-frame consistency (same characters/objects)

You will rank each metric separately for all methods giving 1 to the best method and 4 to the worst. Below are some examples of good/bad generations based on each metric.

Figure 7 presents the detailed explanations of each evaluation criterion, along with visual examples illustrating both high-quality and low-quality generations. These examples were included to calibrate participants’ understanding and ensure consistent evaluation across annotators.

Figure 8 illustrates the evaluation interface and procedure. For each of the five examples, participants were presented with a reference prompt and image, as well as the original story images. They then reviewed the image sequences generated by all four methods, shown in randomized order. For each criterion, participants independently ranked the four methods from 1 (best) to 4 (worst).

Each participant evaluated five examples, resulting in a total of 25 sets of rankings per metric across the study.

Textual Alignment: how well does each frame follow its corresponding prompt. Think about each image separately and look for differences between the textual prompt and the generated image corresponding to that prompt.



Image Alignment: how well does each frame follow the overall style of the reference image. The reference image should be the first frame in the story. The generated images shouldn't steer far away from the style of the reference image. The characters should be similar to the ones introduced as well.



Consistency: how well is the frame-to-frame consistency (same characters/objects) Ignoring the prompts and thinking only about the generated images, are they consistent? Are the characters introduced the same?



Figure 7. Examples illustrating good and poor outputs for each of the three evaluation metrics: textual alignment, image alignment, and consistency. These examples were shown to participants to calibrate their understanding of the evaluation criteria.

D. Limitations

Despite its effectiveness in improving character consistency and text-image alignment, our proposed ViSTA model has certain limitations. First, while our salient history selection strategy enhances efficiency by selecting the most relevant historical reference, it may still struggle in complex narratives where multiple past frames contribute equally to the

Reference Prompt: The city mouse and the country mouse English story.



Original Story

The model doesn't see this, only the previous reference prompt-image



Method 1

Story Prompts:

1. The city mouse went to visit his cousin the country mouse.
2. Country mouse had a humble home.
3. Country Mouse and City Mouse set off together, with the City Mouse inviting his cousin to his grand home.
4. City Mouse proudly announced that a feast awaited us, while the cousins secretly ate delicious foods like ham and chocolate cake.
5. The country mouse decided that the city was not for him.
6. Illustration of a mouse with a basket of fruit.



Methods 2 to 4 are cropped.

Ex1 Rank Textual Alignment (only 1 method per rank) *

	1	2	3	4
Method 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Method 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Method 3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Method 4	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Two other metrics are cropped.

Figure 8. Interface for the human evaluation study. Each example included a reference image and prompt, original story content, and the outputs from four different generation methods (only showing one method for brevity). Participants ranked the methods for each metric independently (only showing one metric for brevity).

current generation. The second limitation is the inherent limitation of auto-regressive methods, error accumulation. As minor imperfections in earlier frames propagate through the sequence, they can compound over time, leading to visual drift, degraded character consistency, or unintended distortions. While our salient history selection strategy mitigates this issue by prioritizing the most relevant historical references, it does not completely eliminate the risk of accumulated inconsistencies. Lastly, our method is evaluated on animated datasets, and its generalizability to photorealistic or highly diverse artistic styles remains an open question.

Reference Image:



Story Prompts:

1. the giraffe , a tall and graceful creature , stands in the grassland , its long neck and legs stretching out to reach the leaves on the tall trees .
2. Giraffe in a classroom, possibly teaching or learning.
3. giraffe in the kitchen
4. The giraffe, with its long neck and legs, is a remarkable creature. It can reach high branches and eat leaves that other animals can't. The giraffe's neck is also very flexible, allowing it to bend down to drink water from a river.
5. The giraffe, a tall and majestic creature, carries a basket of fruits, symbolizing hard work and dedication.
6. In a playful representation of daily life, a giraffe eats a banana in the kitchen, while a monkey drinks a cup of coffee at the table.
7. The giraffe's long neck allows it to reach high branches for food and communicate with other giraffes.

Ground Truth



SDXL-Prompt



IP-Adapter



StoryGen



Ours

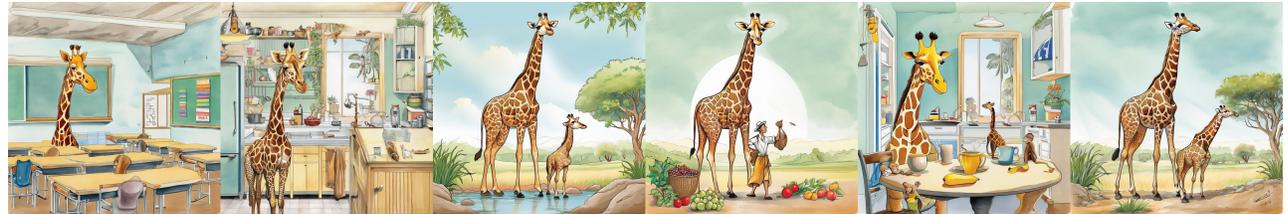


Figure 9. **Additional Qualitative results.** This figure presents a comparison of storytelling between ViSTA, baseline methods, and state-of-the-art on a sample StorySalon story. While SDXL-Prompt show high-quality and well-aligned images, they fail in generating consistent character across all frames. Although IP-Adapter shows consistent character, the generated images do not align with the prompt. Compare with the state-of-the-art StoryGen, our ViSTA shows better consistency on both characters and style.

Reference Image:



Story Prompts:

1. mama khas been harvesting cabbages all week .
2. "Maya and Doobie see a pile of cabbages near Baba Ks old truck. Maya exclaims, 'Wow, there are so many cabbages here!', while Doobie disagrees, saying 'No way, there are only two hundred at the most!'.
3. Mama Kis arrives at the gate and tells everyone to count the cabbages and put them into boxes. There are 20 boxes, and two people can each pack 7 boxes, while one person can pack 6 boxes.
4. children stand around the pile of cabbages talk about different ways to count the cabbages
5. children counting apples
6. children share out the apples equally
7. illustration of a boy and girl cutting an apple

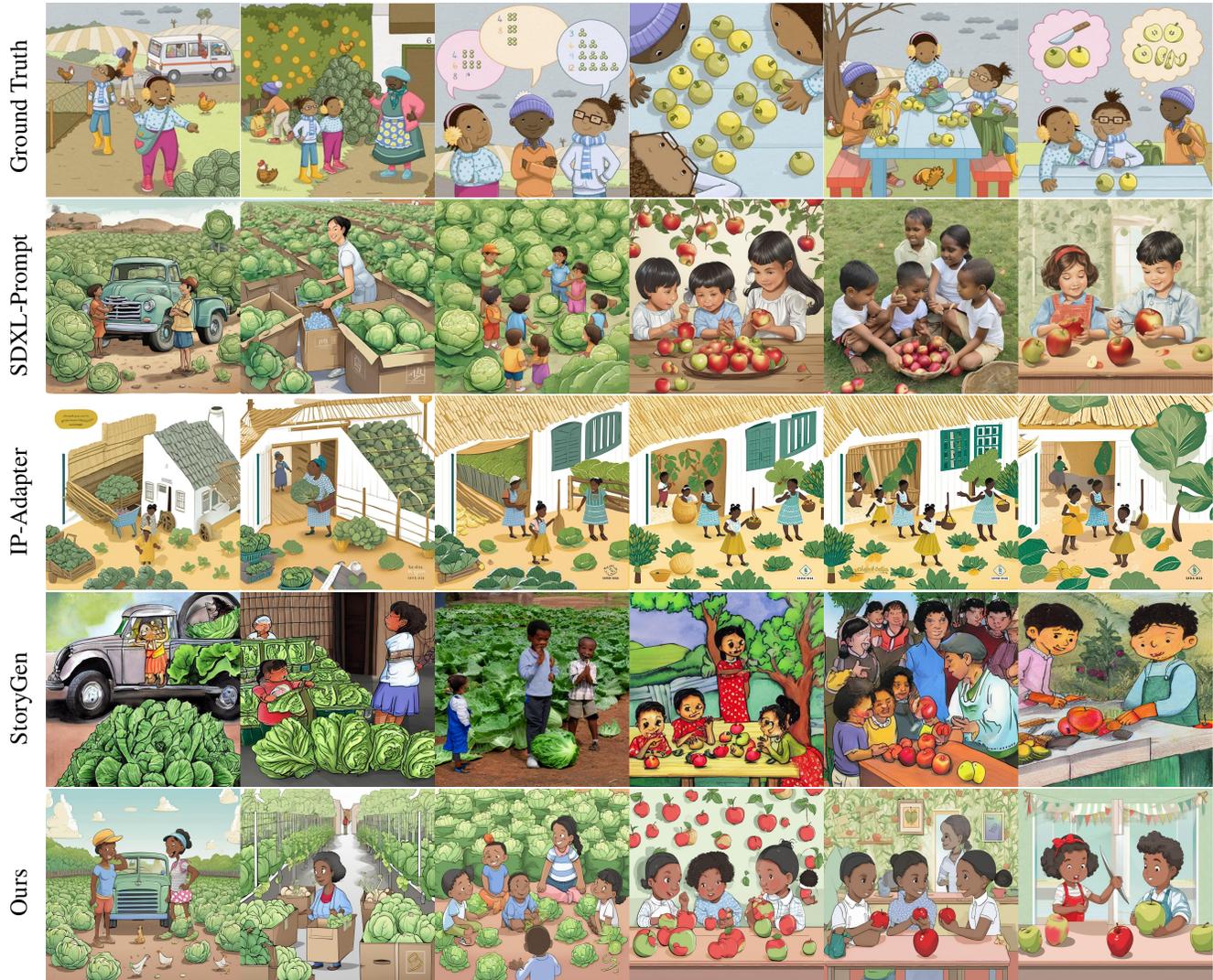


Figure 10. **Additional Qualitative results.** This figure presents a comparison of storytelling between ViSTA, baseline methods, and state-of-the-art on a sample StorySalon story. While SDXL-Prompt show high-quality and well-aligned images, they fail in generating consistent character across all frames. Although IP-Adapter shows consistent character, the generated images do not align with the prompt. Compare with the state-of-the-art StoryGen, our ViSTA shows better consistency on both characters and style.

Reference Image:



Story Prompts:

1. the city mouse and the country mouse english story .
2. the city mouse went to visit his cousin the country mouse
3. country mouse had a humble home .
4. Country Mouse and City Mouse set off together, with the City Mouse inviting his cousin to his grand home.
5. City Mouse proudly announced that a feast awaited USI, while the Cousins secretly ate delicious foods like ham and chocolate cake.
6. the country mouse decided that the city was not for him .
7. illustration of a mouse with a basket of fruit

Ground Truth



SDXL-Prompt



IP-Adapter



StoryGen



Ours



Figure 11. **Additional Qualitative results.** This figure presents a comparison of storytelling between ViSTA, baseline methods, and state-of-the-art on a sample StorySalon story. While SDXL-Prompt show high-quality and well-aligned images, they fail in generating consistent character across all frames. Although IP-Adapter shows consistent character, the generated images do not align with the prompt. Compare with the state-of-the-art StoryGen, our ViSTA shows better consistency on both characters and style.

Reference Image:



Story Prompts:

1. A grade 2 student donated her allowance to help a family in need, inspiring others to do the same, including a teacher who donated his own savings.
2. A boy uses a bamboo rake to harvest rice in a field, symbolizing the hard work and resilience of the people in the rural town of Antique, Sanda.
3. The boy is cooking rice using a traditional method, showing respect for his ancestors. The story is about his dedication to his work and connection to his heritage.
4. The boy in the image is sitting under a tree, possibly in a rural setting. He goes to the forest after completing his work in the house to play with his dog.
5. Children play with a crutch, asking if they can help the boy sitting on the ground.
6. A boy sits under a tree, possibly waiting for someone. The tree symbolizes wisdom and knowledge, and the boy's posture suggests deep thought.
7. A young boy eats a meal with his grandfather, who teaches him proper eating techniques.



Figure 12. **Additional Qualitative results.** This figure presents a comparison of storytelling between ViSTA, baseline methods, and state-of-the-art on a sample StorySalon story. While SDXL-Prompt show high-quality and well-aligned images, they fail in generating consistent character across all frames. Although IP-Adapter shows consistent character, the generated images do not align with the prompt. Compare with the state-of-the-art StoryGen, our ViSTA shows better consistency on both characters and style.