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 SDXL [3] 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 datset, 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

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

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

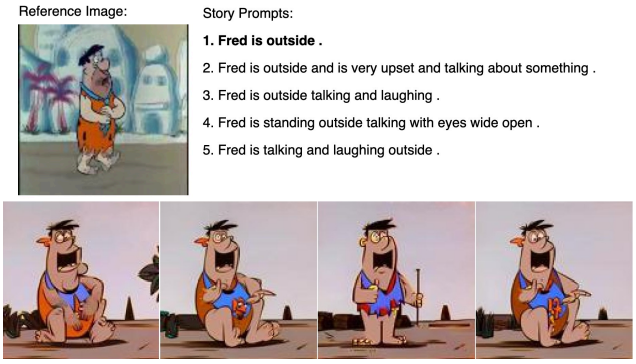

Figure 5. ViSTA sample on FlintStonesSV test set.

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.

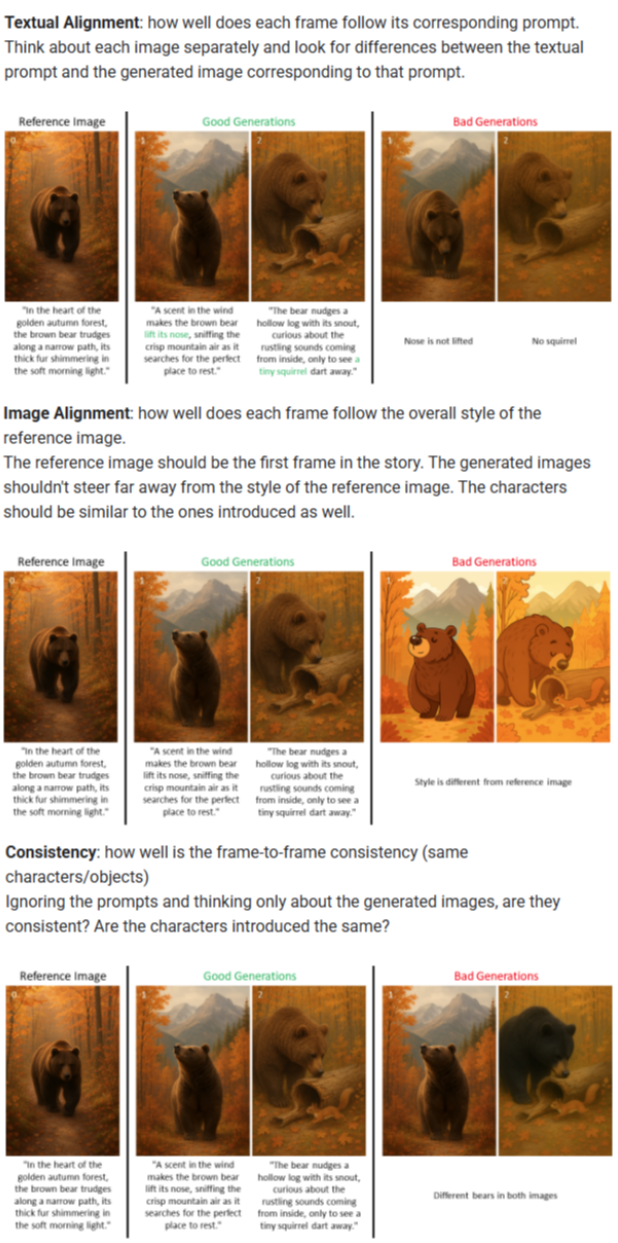

Figure 7. Examples illustrating good and poor outputs for each of the three evaluation metrics: textual alignment, image alignment, and consistency. These examples were shown to participants to calibrate their understanding of the evaluation criteria.

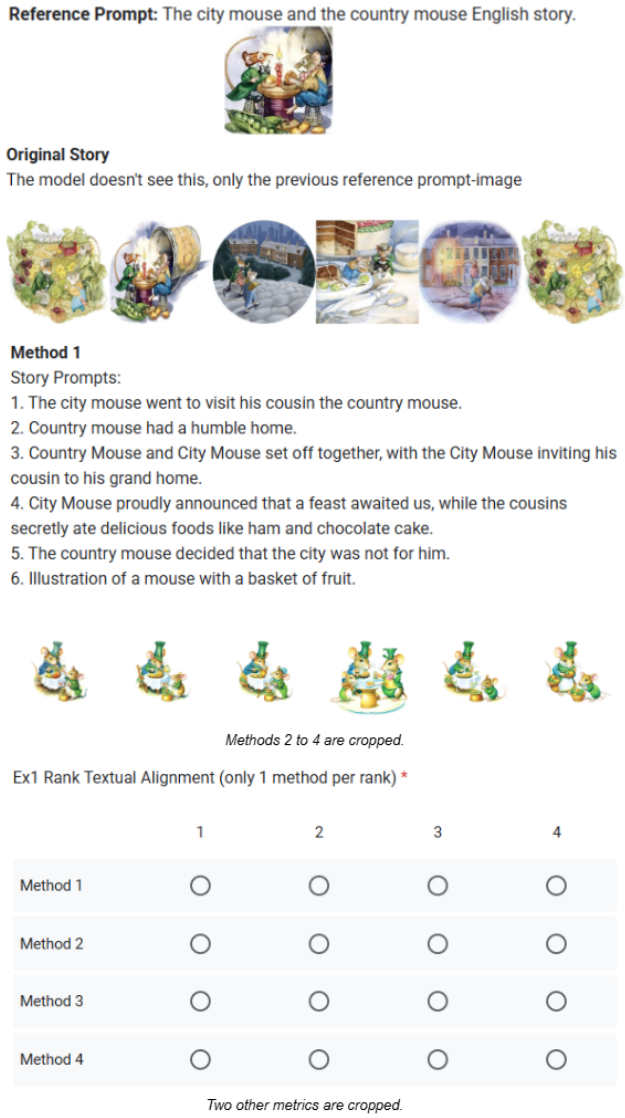

Figure 8. Interface for the human evaluation study. Each example included a reference image and prompt, original story content, and the outputs from four different generation methods (only showing one method for brevity). Participants ranked the methods for each metric independently (only showing one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

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 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.

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 acorss 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.

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 acorss 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.

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 acorss 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.

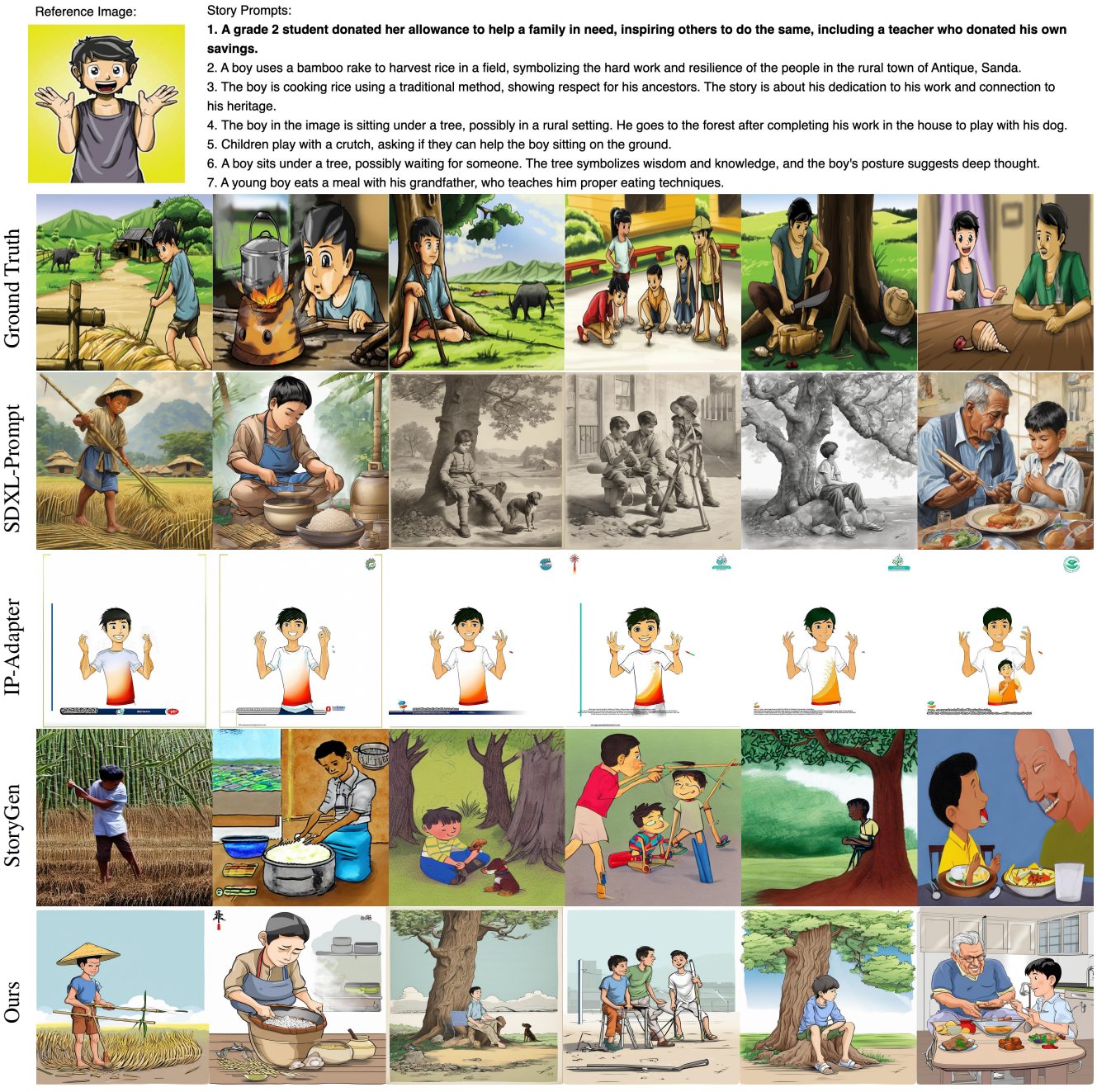

Figure 12. **Additional Qualitative results.** This figure presents a comparison of storytelling between ViSTA, baseline methods, and state-of-the-art on a sample StorySalon story. While SDXL-Prompt show high-quality and well-aligned images, they fail in generating consistent character acorss all frames. 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.