Artificial Intelligence and Social Computing, Vol. 72, 2023, 67–76
https://doi.org/10.54941/ahfe1003275

Automated Visual Story Synthesis with Character Trait Control

Yuetian Chen, Bowen Shi, Peiru Liu, Ruohua Li, and Mei Si

Rensselaer Polytechnic Institute, 110 8th Street, Troy, New York 12180, USA

ABSTRACT

Visual storytelling is an art form that has been utilized for centuries to communicate stories, convey messages, and evoke emotions. The images and text must be used in harmony to create a compelling narrative experience. With the rise of text-to-image generation models such as Stable Diffusion, it is becoming more promising to investigate methods of automatically creating illustrations for stories. However, these diffusion models are usually developed to generate a single image, resulting in a lack of consistency between figures and objects across different illustrations of the same story, which is especially important in stories with human characters. This work introduces a novel technique for creating consistent human figures in visual stories. This is achieved in two steps. The first step is to collect human portraits with various identifying characteristics, such as gender and age, that describe the character. The second step is to use this collection to train DreamBooth to generate a unique token ID for each character prototype. These IDs can then be used to replace the names of the story characters in the image-generation process. By combining these two steps, we can create controlled human figures for various visual storytelling contexts.

Keywords: Visual storytelling, Stable diffusion, Dreambooth, Story synthesis

INTRODUCTION

Visual storytelling is a powerful tool for communication and a rich art form. Visual storytelling provides the opportunity to bring complex ideas and concepts to life. For example, comic strips can effectively communicate a message in a way that is both thought-provoking and accessible. With the advent of automated image and story generation techniques, content creators, educators, and researchers have a wide range of tools at their disposal to produce high-quality visual content with less time and effort.

However, despite these advances, the ability to maintain consistent character appearances throughout a story remains a significant challenge in automated visual story synthesis. Inconsistent character visual traits can compromise the coherence, engagement, and relatability of the story. Characters' appearance and behaviour play a crucial role in shaping the audience's perception of the characters and the plot. Therefore, maintaining consistent visual traits for characters is crucial in creating cohesive, captivating, and relatable stories.

Unfortunately, current diffusion models used for creating story illustrations are designed to produce a single image at a time, which can lead to

inconsistent use of characters and objects across multiple drawings of the same story. This inconsistency can be especially problematic in stories featuring human characters, as the audience's connection to the story often depends on their ability to recognize and relate to the characters.

We have developed a novel solution to tackle the challenge of preserving consistent visual traits in visual storytelling. In our previous work (Chen et al. 2023), we created a story generation pipeline that enables the co-creation of visual stories with users, consisting of two main components: narrative and image generation. The narrative generation empowers the user to control the events and emotions in the generated content, while the image generation produces corresponding illustrations through a diffusion model.

Our current work builds upon the previous pipeline by focusing on enhancing the consistency of the generated images. We achieve this by fine-tuning the Stable Diffusion model through Dream- Booth (Ruiz et al. 2022) and the VGGFace2 HQ dataset (Cao et al. 2017). This approach enables us to produce coherent and consistent illustrations for visual storytelling while preserving the unique visual traits of the characters.

As illustrated in Figure 1, our model can generate illustrations for a short story that features a specific character, with the help of a reference photo. To assist users in selecting an appropriate reference photo for their story characters, we have developed a keyword-based approach. In Section EXAMPLE OUTPUT, we present various examples of story drawings that exhibit consistent and identifiable characters, demonstrating the effectiveness of our approach. We also provide a comprehensive analysis of the limitations and potential areas for future research.

RELATED WORK

We will discuss related research in two areas related to visual story co-creation: automatic story generation and image generation. Large-scale

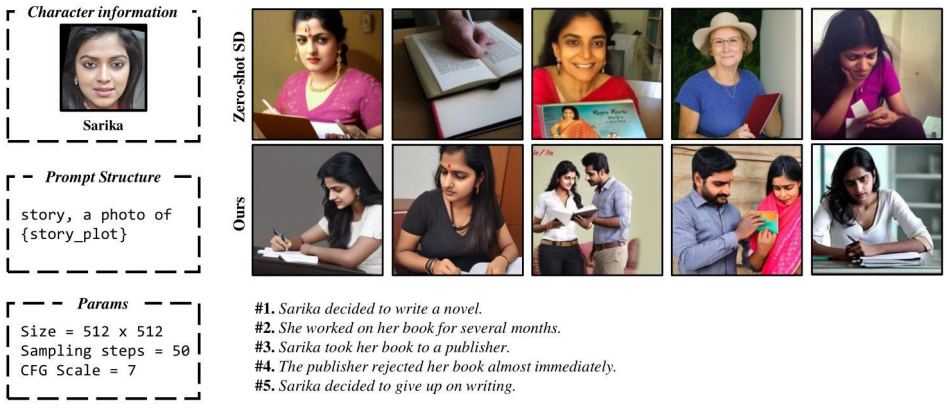

Figure 1: Comparison of a visual story example synthesized by zero-shot stable diffusion and our algorithm.

language models have been shown to be effective in producing cohesive stories, while diffusion models have become popular for generating realistic images. Additionally, maintaining consistency among the generated images has been the subject of extensive research, providing insights for developing high-quality visual stories. The DreamBooth (Ruiz et al. 2022) technique, in particular, serves as the foundation for our proposed method for producing consistent character visualizations.

Automatic Visual Story Generation

In recent years, Transformer-based models, and their variants, such as Bert (Devlin et al. 2018), GPT- 3 (Brown et al. 2020), and Chat-GPT, have emerged as powerful tools for automatic story generation and natural language processing tasks. These pre-trained large language models have demonstrated unprecedented generative capabilities, allowing for the production of coherent and imaginative stories with relative ease. While automatic story generation has many potential applications, including education and entertainment, the addition of visual elements may always improve the overall narrative experience.

In our previous work (Chen et al. 2023), we developed a visual story co-creation pipeline with two main components: the next sentence generation module and the image generation module. The next sentence generation module creates a story sentence by sentence based on user-specified keywords and character emotions. The generated sentence is then visualized by the image generation module. In our previous work, the workflow of our visual storytelling pipeline is iterative, with the system suggesting characters' emotions and words for the next sentence based on the partial story already written. The user can either accept the system's recommendations or provide their own. The user is presented with several visual representations of the new sentence and chooses one to proceed with. Finally, object detection is applied to the generated images in order to extract additional keywords for future co-creation processes.

While we achieved satisfactory results in sentence generation with the given keywords and emotions, the consistency of the generated images in our previous work was lacking. In a baseline case, the main character in the generated images varies significantly, which hinders the narrative experience. As a result, in our current work, we focus on improving image consistency by fine-tuning the Stable Diffusion model using the DreamBooth technique (Ruiz et al. 2022) and the VGGFace2 HQ dataset (Cao et al. 2017) using a keyword-based approach.

Text-to-Image Synthesis

The field of generating images from textual descriptions has made significant progress in visual realism, diversity, and semantic alignment. Large-scale diffusion models (Ramesh et al. 2021) have recently shown impressive results in generating images from textual input. These models, such as the Stable Diffusion model (Rombach et al. 2021), rely on large datasets of annotated image-text pairs for training and can synthesize high-quality images with

controllable attributes. The Stable Diffusion model can generate a diverse range of visuals, including those with human figures, and is a promising candidate for integration into automated visual story-generating pipelines as a text-to-image synthesizer.

In previous work, we explored incorporating the Stable Diffusion model into a visual story generation pipeline. However, diffusion models are designed to produce a single image at a time, which poses challenges for generating a series of images with consistent characters and objects. Additionally, detailed descriptions are necessary to instruct the model to create characters with specific attributes in an image, such as "a young girl with long red hair and pale skin." Without such descriptions, the generated characters will not remain consistent throughout the story. Several approaches have been developed to address the challenge of maintaining the consistency of characters and objects across successive generated images in text-to-image models. For example, one approach (Dhariwal and Nichol 2021) allows the user to provide a mask to constrain changes to specific regions of the image, while another (Hertz et al. 2022) enables editing without using masks. Textual Inversion (Gal et al. 2022a) proposes using pseudo-words in the embedding space of a pre-trained text-to-image model to guide the generation process. AR-LDM (Pan et al. 2022) is another approach that uses a latent diffusion model trained on historical captions and generated images to produce a series of consistent visuals for storytelling. In our work, we propose a system that employs a different technology to achieve a similar goal. Specifically, we utilize Dreambooth, which has shown to be reliable and accurate in producing images with consistent objects and figures (Ruiz et al. 2022).

DreamBooth

DreamBooth is a powerful tool that allows users to fine-tune large text-to-image generative models to learn new concepts. This tool utilizes a few reference images to fine-tune the diffusion model to learn a new concept, such as a male figure, which is represented by a special token. This token can be used in conjunction with other descriptions of the man, such as his mood, behavior, attire, and facial expressions, to generate images of the man in various situations, positions, perspectives, and lighting conditions. To specify the explicit features of the characters in their stories, we provide a keyword-based strategy.

When fine-tuning all layers of the diffusion model with only a few sample images of the subject, there is a risk of experiencing language drift, where prior knowledge is lost during the few-shot fine-tuning process as the layers holding the prior knowledge are altered during training. This can result in a decline in the quality of the generated images due to a shift in the model's language generation. To mitigate this issue, DreamBooth employs an autogenous class-specific prior-preserving loss, as shown in Equation 1. This loss supervises the model with its own generated images, encouraging the model to maintain its prior knowledge throughout the image generation process.

$$E_{x,c,\epsilon,\epsilon',t}\left[w_t \left| |\widehat{x_\theta}\left(\alpha_t \mathbf{x} + \sigma_t \epsilon, c\right)| \right|_2^2 + \lambda w_{t'} \left| \left|\widehat{x_\theta}\left(\alpha_{t'} \mathbf{x}_{\mathrm{pr}} + \sigma_{t'} \epsilon', c_{\mathrm{pr}}\right) - x_{\mathrm{pr}}\right| \right|_2^2\right] (1)$$

PROPOSED APPROACH

Our approach to creating consistent character visualization involves using real human portraits as prototypes. The challenge is to map a story character to one of these prototypes since detailed character descriptions are not always included in a story. For instance, in Figure 1, it is natural to simply say, "Sarika is a dedicated writer..." without a detailed physical description.

To address this issue, we built a character repository with a variety of attributes that can be used to index these characters. Using these character prototypes, we then fine-tuned a Stable Diffusion model. Users can access this repository during the visual story creation process and make selections according to their desired character attributes as shown in Figure 3. The characters chosen by the users will be used in the story's illustrations. Currently, our system includes nine-character prototypes and two attributes, age range, and gender. The characters are labeled using a convention of <gender, age, id>. Two examples of this convention are <male, mid-aged, 01> and <male, mid-aged, 02>, which represent two different prototypes of a mid-aged male character.

In order to implement our proposed approach, a portrait dataset is required for each character prototype, which is then used as input to a neural network-based facial recognition system to extract facial attributes. These features are then used to identify the character prototype. Figure 2 illustrates the steps of our approach:

- Collect a portrait dataset for each of the 9 different character prototypes.
- Use a neural network-based facial recognition system to extract facial attributes, including age range and gender, for each character prototype.

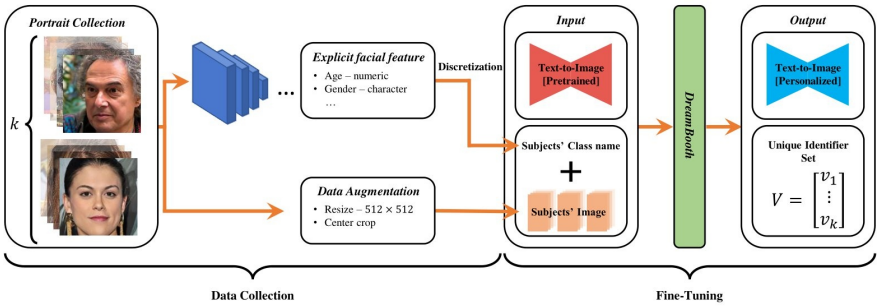

Figure 2: Pipeline for finetuning.

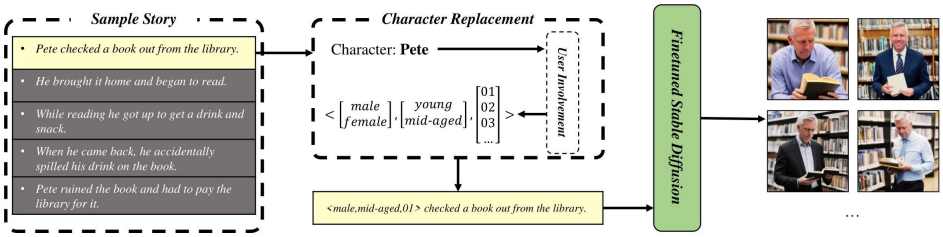

Figure 3: The pipeline at the deployment stage.

- Fine-tune the Stable Diffusion model using DreamBooth for each character prototype.

Portrait Dataset Aggregation

To establish a robust base of image classes to fine-tune the Stable Diffusion model, we need a significant amount of data. For this purpose, the VGGFace2 HQ dataset (Cao et al. 2017) provides a valuable resource with over 3 million high-quality face images that can be used as prototypes for human characters in visual stories. The dataset offers a diverse set of attributes, including age, gender, expression, pose, and race, which adds to the variety of potential protagonists in visual stories. We selected 9 character prototypes – 9 classes, each representing a human character that varies with facial attributes, from this dataset. As we have 400 high-quality images for each of the 9 classes, we can ensure the robustness of the generation by utilizing the portrait images during the fine-tuning process.

Facial Attribute Extraction

The VGGFace2 HQ dataset doesn't include annotations for images, such as gender or age, which makes it difficult to index prototypes for users to choose from. While manual annotation is one way to generate such annotations, it is impractical due to the large number of images in the dataset. As an alternative approach, we used machine learning to extract information about gender and age automatically. Specifically, we used a classifier provided by (Kim 2021), which can identify gender and age in the format of a tuple where the first element is gender and the second one is float representing the approximate age.

Finetuning Stable Diffusion

To fine-tune the Stable Diffusion model with DreamBooth, we use a set of selected face images from the VGGFace2 HQ dataset, denoted by S. Let s_n be the n-th human character in S, where $n = 1, 2,..., |S|$. The model parameters are denoted by ω. In order to perform multi-token finetuning with DreamBooth, we create a character prototype set C_n, where we have.

$$C_n = \begin{bmatrix} s_1 \\ \vdots \\ s_k \end{bmatrix} \quad and\ k \leq n$$

During the finetuning process, 50 images are randomly selected from s_n to serve as a character prototype c_n, and these prototypes are combined with conceptual knowledge from the prompt "a photo of human" to encourage generalizability. The objective of DreamBooth is to minimize the overall loss across all target images in S, which can be expressed as:

$$\sum_{k=1}^{||C_n||} argmin_{\omega} \frac{1}{||S_n||} \sum_{k=1}^{||C_n||} \mathrm{L}(\omega, s_i) \tag{2}$$

For instance, if S contains 10-character prototypes, then the objective of equation (2) is to minimize the singular loss across all 10 prototypes in each

iteration. This process of updating the model parameters using guidance from the target images continues until the model converges or a stopping criterion is met.

Example Output

Multiple examples of automated visual story generation using our finetuned Stable Diffusion model that prioritizes facial attribute consistency is presented in Figure 4. The stories were selected from the ROCStories dataset, which is a well-established collection of short five-sentence stories (Mostafazadeh et al. 2016).

Our model is based on the Stable Diffusion v1.5 architecture, and we fine-tune it by adjusting the number of training steps to balance the detail of the features and robustness of the model in deployment. We observed that increasing the number of steps resulted in more detailed generated images but also led to an increase in stereotypical features, such as the loss of skin detail and the intensification of wrinkles. To strike a balance between accuracy and naturalness, we fine-tuned the model for 850 steps for each of the 9 classes. This process took approximately 35 minutes per class on a single NVIDIA Tesla T4-16GB GPU.

RESULTS AND DISCUSSION

Our approach has been successful in preserving important visual features of the characters in their stories, as demonstrated by our results in Figure 4. In each case, we compare the story visualizations generated by our approach with those generated using the Stable Diffusion model without any fine-tuning (referred to as Zero-shot SD).

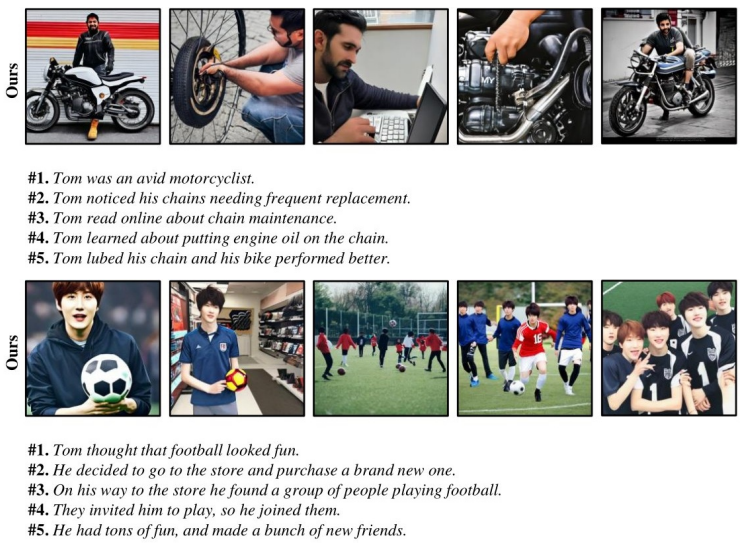

Figure 4: Visual story synthesis results.

The first example in Figure 4 features a character with a distinctive beard. Despite appearing in different poses and expressions, the character's beard remains unchanged and consistent throughout the story. This highlights the effectiveness of our approach in preserving the unique features of the character and ensuring their recognizability and memorability throughout the narrative. Similarly, in the second example, we showcase the preservation of ethnographic features, such as skin color and facial features, that remain consistent throughout the story. Also, Figure 5 shows how the same story can be visually represented with two different characters. In both cases, the visualization appears to be quite natural. As a result, our system allows authors to experiment with different character designs and personalize their stories based on their preferences.

Although our approach has shown promising results, there are still several limitations that the im- age generation model faces, as highlighted in Figure 6. These limitations can be broadly categorized into two main issues: subject confusion, and prototype confusion.

Subject confusion arises from the model's limited prior knowledge of certain subjects. This can result in the model's inability to understand certain activities of characters, such as recognizing actions like "thought," "believe," and "want." For instance, if the prompt is "a person thinking reading is hard," the model may struggle to generate an image of a person thinking as if it is

Figure 5: Visual story synthesis results with different characters.

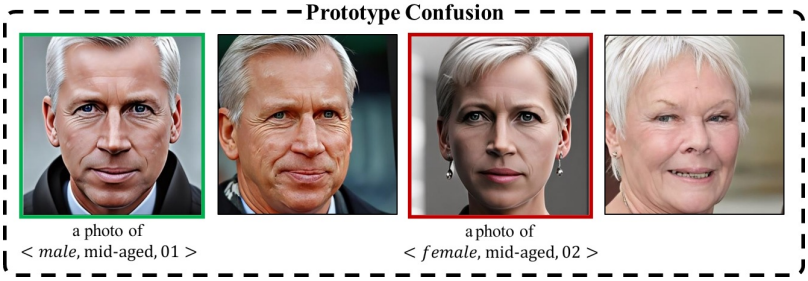

Figure 6: Unexpected outcomes of our image generation method.

an abstract action. This limitation stems from the model's lack of prior knowledge about the nuances of human behavior and can result in inaccurate or unrealistic generated images.

Prototype confusion occurs when the model confuses one character's prototype with another due to similarities in their names. For instance, in the second example shown in Figure 6, the names <female, mid-aged, 02> and <male, mid-aged, 01> are very similar to each other, leading to confusion in the model. As a result, the features of the latter, such as the shape of the eye sockets and nose wings, interfere with the generation of the former, causing an increase in the disparity between the generated image and the ground truth. This interference can lead to a failure to control the outcome, which is consistent with the findings in the textual inversion work (Gal et al. 2022b).

Future Work

The combination of automatic story and image generation has great potential for creating engaging and expressive stories, but there is still room for improvement. One potential improvement is incorporating metrics or measuring methods to evaluate the training performance, as the current approach lacks ground truth for testing generated images. Second, the VGGFace2 HQ dataset is useful for training face-generation models, but it may have biases that could affect how well the model works. For example, since all the faces are of celebrities, it might be hard to generalize to the general public. The dataset is also biased toward people between the ages of 25 and 31, which makes it harder to generalize to people younger or older. Due to the unequal distribution of different races, ethnicity can also be a source of bias. Also, since the dataset only has pictures of real people, it might be hard for the model to make pictures that are artistic, like cartoons. To reduce these biases, future work could combine data from different sources, such as artistically styled images. Additionally, the model's controlled features could be expanded to include other facial attributes, such as eye color, expression, poses, or even other objects that appeared in the story. Finally, the model's robustness and generalizability can be improved with the use of techniques like adversarial training or data augmentation.

CONCLUSION

This paper presents a novel approach to visual story co-creation, which aims to bridge the gap be- tween automatic story generation and visual storytelling. The proposed system harnesses the power of large language models and deep learning to create short visual stories in collaboration with an author. Further, by fine-tuning Stable Diffusion via DreamBooth the system is able to create unique tokens for different character prototypes. By using labels to index these prototypes, users are granted more control over the appearances of the generated characters while maintaining their visual consistency across multiple images.

REFERENCES

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.;

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2017. Vggface2: A dataset for recognising faces across pose and age.

Chen, Y.; Li, R.; Shi, B.; Liu, P.; and Si, M. 2023. Visual story generation based on emotion and keywords.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Dhariwal, P., and Nichol, A. 2021. Diffusion models beat gans on image synthesis.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022a. An image is worth one word: Personalizing text-to-image generation using textual inversion.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022b. An image is worth one word: Personalizing text-to-image generation using textual inversion.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control.

Kim, T. 2021. Generalizing mlps with dropouts, batch normalization, and skip connections. arXiv preprint arXiv:2108.08186.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories.

Pan, X.; Qin, P.; Li, Y.; Xue, H.; and Chen, W. 2022. Synthesizing coherent story with autoregressive latent diffusion models.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation.

Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-resolution image synthesis with latent diffusion models.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.