

Supplementary Materials of Evolving Storytelling: Benchmarks and Methods for New Character Customization with Diffusion Models

Anonymous Authors

1 INTRODUCTION

This supplementary material presents more detail about our paper Evolving Storytelling: Benchmarks and Methods for New Character Customization with Diffusion Models. We organized this supplementary as follows: 1) we present more details regarding the datasets contained in NewEpisode; 2) we present detailed experiment results and a user study regarding the generated images of EpicEvo on NewEpisode_{Flintstones} and NewEpisode_{Pororo}; 3) we present more visual examples generated by EpicEvo; 4) we discuss the limitations of our method; and finally 6) we discuss potential societal impacts.

2 DATASET DETAILS

In this section, we introduce the details about our benchmark NewEpisode. Specifically, it contains two customization benchmarks, namely NewEpisode_{Flintstones} and NewEpisode_{Pororo}, derived from the original FlintstonesSV [8] and PororoSV [6] datasets. The main characters in NewEpisode_{Flintstones} are shown in Fig. 1.

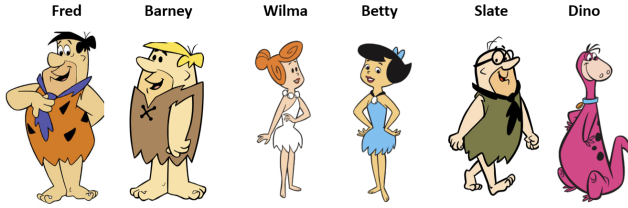


Figure 1: Main characters of the NewEpisode_{Flintstones}

The new characters available for the story character customization task are shown in Fig. 2.



Figure 2: New characters of the NewEpisode_{Flintstones}

For NewEpisode_{Pororo}, the main characters are shown in Fig. 3, and the new characters are shown in Fig. 4.

For each new character, its sample size is shown in Table. 2. We list both the available images and available stories for each



Figure 3: Main characters of the NewEpisode_{Pororo}



Figure 4: New characters of the NewEpisode_{Pororo}

character. Notably, some images can reoccur in other stories due to the original construction process of FlintstonesSV [8], *i.e.*, 5 logically consecutive images are considered one story. During the evaluation, we evaluate all the corresponding stories and calculate the FID [2] scores for each character for NewEpisode_{Flintstones}. We randomly select a maximum number of 100 stories (with a fixed random seed) for NewEpisode_{Pororo} and evaluate all images related to the new characters. We report the average FID scores and CLIP-based scores [10] to avoid the case that characters who have many more stories, *e.g.*, Gazoo, can overshadow the model performance on characters that have lesser evaluation samples. For model customization, we use one story of each character to finetune the model. Notably, we use stories that contain images that are all related to the new character because this leads to more efficient customization, and is closer to real-world scenarios (it is easier for the end user to provide images that only contain the new character). Therefore, during evaluation, these 5 images used for customization will not be included. The samples used for customization are fixed across all our baselines. Code and dataset will be released.

3 EXTENDED MODEL EVALUATION

In this section, we provide the extended evaluation results of our story character customization model, EpicEvo, and our baselines,

Table 1: Detailed FID scores for each new character in NewEpisode

Characters	NewEpisode FID			
	Textual Inversion [1]	DreamBooth [12]	Custom Diffusion [5]	EpicEvo(Ours)
Slaghoople	150.21	139.68	136.47	130.08
Tex Hardrock v1	250.05	222.26	233.73	228.12
Gazoo	127.94	114.82	120.97	118.91
Police in Helmet	234.41	223.91	206.30	189.46
Pianist	261.59	285.26	227.67	229.93
Rockzilla	230.22	205.57	220.82	206.39
Tex Hardrock v2	179.51	160.79	173.40	166.38
Theft	195.93	164.03	175.35	182.11
Seal	236.39	220.26	224.04	243.30
Popo	144.19	138.41	135.33	140.15
Pipi	139.80	138.26	133.66	137.06
Whale	202.52	154.02	162.69	146.75
Shark	148.21	128.52	114.04	116.81
Harry	112.80	116.07	118.13	115.01
Tutu	152.92	134.33	127.39	126.60

Table 2: Dataset details of NewEpisode

	Available Images	Available Stories
Slaghoople	73	139
Tex Hardrock v1	19	22
Gazoo	138	268
Police in Helmet	19	29
Pianist	23	33
Rockzilla	26	37
Tex Hardrock v2	41	72
Theft	21	34
Seal	23	44
Tutu	184	331
Popo	129	186
Pipi	120	187
Shark	150	245
Whale	17	36
Harry	1980	3714

i.e., DreamBooth [12], Custom Diffusion [5], and Textual Inversion [1]. We first show the detailed FID score for each new character in NewEpisode. We present the results for all 15 new characters in Table 1. To further study the model performance, we also conducted a user preference study as some works like [9] suggested that automatic metrics such as FID [2] might not fully align with human perception. Specifically, we present a comparison-based user study where each user is prompted to pick the image that best matches the ground truth image and the input text prompt, visual samples in Fig. 5,6,7 are some of the samples we used during the user study. Each user is allowed to select at most 2 matching images if there is no obvious winner, the users are also allowed to skip the current comparison if they think there are no matching results. We collected 20 user studies from users with various backgrounds and summarized the user preference rate in Table. 3. Note that we mark the case where users skip the comparison as 'Tie', meaning all generated images are somewhat uncorrelated with the ground truth and/or the text prompt. In conclusion, we found that the FID score alone might be insufficient to reflect the actual synthesis quality as we found that a smaller FID score difference does not deterministically lead to worse or better visual quality. For instance, in the case of *Tex Hardrock v1*, we found our model is more frequently preferred despite it does not reach the lowest FID. Similarly, for the *Shark* and the *Whale*, we found [12] is more

Table 3: User preference rate for each character across all available characters in NewEpisode^{Flintstones} and NewEpisode^{Pororo}

	Tie	TI [1]	DB [12]	CustomDiff [5]	EpicEvo
Slaghoople	7.09%	3.94%	30.71%	23.62%	34.65%
Tex Hardrock v1	12.22%	1.11%	21.11%	7.78%	57.78%
Gazoo	9.50%	1.81%	37.10%	8.14%	43.44%
Police in Helmet	7.69%	9.89%	25.27%	9.89%	47.25%
Pianist	1.10%	1.10%	14.29%	20.88%	62.64%
Rockzilla	4.47%	0.00%	39.11%	2.23%	54.19%
Tex Hardrock v2	12.08%	0.00%	14.77%	4.70%	68.46%
Theft	10.47%	0.00%	24.42%	36.05%	29.07%
Seal	27.08%	4.17%	36.46%	16.67%	15.62%
Tutu	24.64%	16.30%	14.13%	10.51%	34.42%
Popo	15.70%	3.31%	42.98%	10.33%	27.69%
Pipi	15.73%	2.25%	29.29%	13.86%	39.58%
Shark	18.81%	16.83%	40.59%	8.91%	14.85%
Whale	15.38%	0.00%	51.28%	24.36%	8.97%
Harry	20.77%	10.38%	19.67%	22.95%	26.23%

frequently preferred despite having a higher FID score. Still, there are also cases in which FID could lead to conclusions about model performances, *e.g.*, *Slaghoople*, *Police in Helmet*, *Pianist*, *Tutu*, and *Harry*.

We hypothesized that the FID score could ignore aspects including the level of artifacts, texture quality, and subtle semantic details while attending to aspects deemed less important by users such as background and style. Additionally, our preliminary studies also show that the current vision-language model also shows an unsatisfactory level of accuracy as their visual encoder could ignore subtle details [7]. In conclusion, based on the user study, we found EpicEvo winning in 66.67% of cases, [12] winning in 26.67% of cases, and [5] winning in 6.67% of cases. We found [12], despite being simple, is more effective than [5]. This is contrary to the conclusions made by [5], *i.e.*, tuning only the cross-attention layer could result in better multi-concept generation performance. We hypothesize that such a conclusion might be limited to the case of tuning a general text-to-image model such as Stable Diffusion [11]. In the case of the model specialized for storytelling, tuning the cross-attention layers might lead to the underfitting of customization samples. We also found using LoRA [3] also leads to unsatisfactory results, and



Figure 5: Visual examples of model-generated stories for new and existing characters.

this might also be attributed to the fact that an insufficient number of parameters could underfit the customization samples.

In sum, in this section, we discuss the customized model performance more thoroughly using automatic metrics and human evaluation results. We found that our model could generate less satisfactory results for certain new characters despite achieving better automatic metric scores. This suggests that it is necessary to

consider evaluation results from various aspects to correctly assess model performance.

4 QUALITATIVE MODEL EVALUATION

In this section, we present more visual illustrations of model-generated results from our method and the baselines.



Figure 6: Visual examples of model-generated stories for new and existing characters.

In Fig. 5, stories related to *Slaghoople*, *Tex Hardrock v1*, *Gazoo*, *Pianist*, *Police in Helmet*, and *Rockzilla* are displayed. We found that our method is more capable of generating images that contains multiple characters. For instance, in (11), (15), and (16) EpicEvo could correctly generate the mentioned characters in the text input, demonstrating a stronger ability to synthesize stories with more complex dynamics. In the case of (1) and (8), we also found that our

method could generate stories that have better character consistency, *i.e.*, the generated character is closer to the new character we customized. Admittedly, there are failure cases such as (4) where our model seems to be overwhelmingly affected by the presence of *Slaghoople* in the input text, leading to the result of generating two characters that have her appearance.



Figure 7: Visual examples of model-generated stories for new and existing characters.

In this figure, we show examples related to *Rockzilla*, *Tex Hardrock v2*, *Seal*, and *Theft*. We found that our model is more performing on generating characters such as *Rockzilla* as we observe the generated characters could better illustrate actions between *Rockzilla* and existing characters such as *Barney*. Additionally, EpicEvo also synthesizes stories for *Tex Hardrock v2* better than other competitive methods by depicting the character more precisely, albeit with some

level of misalignment remaining. Nevertheless, we validate the case in our user study that EpicEvo is less capable of generating stories for the *Theft* and the *Seal*. Overall, our method empirically performs better than the baselines on *NewEpisode_{Flintstoens}*. We also display results from *NewEpisode_{Pororo}* in this figure. We found a noticeable improvement in stories containing the character *Tutu*

as the existing characters appear more accurately. For images related to Popo and Pipi, we found that although our method could distinguish more between these two very similar characters, the adversarial training procedure could lead to more artifacts, leading to unsatisfactory results. This is also reflected by our user study, as we have a similar preference rate as DreamBooth [12].

In Fig. 7, we continue to display samples for the *Shark*, the *Whale*, and *Harry*. As indicated by the user study, we found that DreamBooth [12] captures the characteristics of the *Shark* and the *Whale* more accurately, leading to a higher user preference rate. Still, we empirically found this might be a result of overfitting. For instance, in Fig. 7-(9), DreamBooth [12] seems to be overfitted to the customization samples as the generated image is less aligned with the caption despite correctly depicting the new character. Lastly, we discuss the case of *Harry*. *Harry* is a rather frequent character in the original dataset and it proves to be very challenging to generate for all customized models. We hypothesize the reason could be that *Harry* is significantly smaller than most other existing characters. With compression processes like VAE [4], it could be challenging for the model to correctly learn the representation for such characters and this leads to unsatisfactory results. Nonetheless, our adversarial character alignment method seems to be able to encourage the generating of such novel characters, despite having plenty of room for improvement.

5 LIMITATIONS

In conclusion, we introduced EpicEvo, a suite of methods tailored for story character customization. This specific application presents unique challenges not typically encountered in standard customization tasks. For instance, the model inherently carries strong priors about existing characters, which can complicate the introduction of new characters. Our approach incorporates an adversarial character alignment module aimed at fostering the generation of narratives for these new characters. Despite our efforts, our detailed evaluations reveal that EpicEvo has not fully succeeded in generating coherent stories for new characters. Additionally, we observed a tendency for the story generation model to overfit the customization examples. Moreover, the adversarial nature of the character alignment process sometimes introduces visual artifacts in the generated outputs, a consequence of the inherent instability in adversarial training methods. Addressing this instability remains a task for future research. Furthermore, our visual assessments highlighted the persistent challenges in generating complex compositions, such as images featuring multiple characters. These difficulties are not exclusive to our method but are also prevalent in diffusion models and other diffusion-based story generation approaches. Exploring new techniques to better control and regulate character representation in narrative synthesis could be a valuable direction for future research, potentially enhancing fidelity to textual descriptions in generated visual content.

6 SOCIETAL IMPACT

The development of story character customization methods like EpicEvo offers significant creative potential and economic opportunities by enabling more diverse and personalized storytelling.

However, these technologies also pose risks, such as the potential for misuse in creating harmful or misleading content. This raises important ethical and legal challenges, particularly concerning copyright issues and the propagation of stereotypes. Balancing these benefits and risks will require careful management, including the development of ethical guidelines and robust legal frameworks to govern the use of such.

REFERENCES

- [1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. <https://doi.org/10.48550/ARXIV.2208.01618>
- [2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.), 6626–6637. <https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html>
- [3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=nZeVKeeFYf9>
- [4] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1312.6114>
- [5] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 1931–1941. <https://doi.org/10.1109/CVPR52729.2023.00192>
- [6] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuxin Wu, Lawrence Carin, David E. Carlson, and Jianfeng Gao. 2019. StoryGAN: A Sequential Conditional GAN for Story Visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 6329–6338. <https://doi.org/10.1109/CVPR.2019.00649>
- [7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.
- [8] Adyasha Maharana and Mohit Bansal. 2021. Integrating Visuospatial, Linguistic, and Commonsense Structure into Story Visualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 6772–6786. <https://doi.org/10.18653/1/2021.EMNLP-MAIN.543>
- [9] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhui Chen. 2022. Synthesizing Coherent Story with Auto-Regressive Latent Diffusion Models. *CoRR* abs/2211.10950 (2022). <https://doi.org/10.48550/ARXIV.2211.10950>
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752 [cs.CV]*
- [12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 22500–22510. <https://doi.org/10.1109/CVPR52729.2023.02155>