Lay2Story: Extending Diffusion Transformers for Layout-Togglable Story Generation

Supplementary Material

A. Related Work

A.1. Consistent Text-to-image Generation

Consistent image generation methods can be categorized into high-level semantic consistency, facial consistency, style consistency, and object consistency [36]. High-level semantic consistency methods [10, 11, 16, 39], such as ReVersion [11], achieve consistency by inverting object relations and utilizing a contrastive loss to guide the optimization of token embeddings toward specific clusters of Part-of-Speech tags, such as prepositions, nouns, and verbs. Facial consistency methods [12, 25, 31, 40], such as PhotoMaker [12], construct high-quality datasets through meticulous data collection and filtering pipelines, employing a two-layer MLP to fuse ID features and class embeddings for comprehensive human portrait representation. Style consistency methods [8, 22, 29, 37], such as StyleAdapter [29], introduce a specialized embedding module to extract and integrate global features from multiple style references and employ a dual-path cross-attention mechanism within the learning framework. Object consistency methods [4, 27, 28] include approaches like IP-Adapter [34], which trains a lightweight decoupled cross-attention module where image and text features are processed separately with query features; DreamBooth [21], which proposes using a unique modifier with a rare token to represent the subject of interest and fine-tuning all parameters of the diffusion model; and UMM-Diffusion [14], which designs a multi-modal encoder to generate fused features based on the reference image and text prompt. Storytelling task can essentially be categorized as an object consistency image generation task, aiming to achieve consistent visual narratives through cross-modal fusion [36].

A.2. Layout-to-image Generation

Layout-controllable image generation aims to apply layout control to place subjects in user-defined positions within an image, which has become an active research area [7, 15, 30, 38]. SimM [5] is a training-free system that corrects layout errors during inference by analyzing prompts, detecting inconsistencies, and adjusting activations. ReCo [33] introduces a unified token vocabulary containing both text and positional tokens for precise, open-ended regional control. InteractDiffusion [9] enhances T2I diffusion models by incorporating Human-Object Interaction (HOI) information through tokenized embeddings and a self-attention layer, enabling better control of interactions and locations in generated images. CreatiLayout [35] introduces a Siamese architecture to decouple image-layout interactions in MM-DiT, treating layout as an independent modality and integrating it with text and image features while leveraging a large-scale dataset for training and evaluation. Combining the Layout-to-Image task with the Storytelling task is both innovative and valuable.

A.3. Storytelling Generation

Generating a sequence of frames with a consistent subject from a given script, known as storytelling, is a rapidly evolving field. Current methods are generally categorized into two types: training-free and training-based. Training-free methods, such as StoryDiffusion [41], utilize consistent self-attention computation based on the SD1.5 [19] model to maintain subject consistency throughout the story sequence. ConsiStory [24] achieves subject consistency by sharing the internal activations of the pre-trained diffusion model. 1Prompt1Story [13] takes advantage of the inherent context consistency of language models, using a single prompt to generate a cohesive narrative across the story sequence. Training-based methods, such as Seed-Story [32], employ the Multimodal Large Language Model (MLLM) to predict text and visual tokens, followed by a visual de-tokenizer to ensure subject consistency across the image sequence. FLUX.1-dev IP-Adapter [23] builds upon the robust image generation model FLUX [1], training an adapter to integrate reference image features, enabling FLUX to generate images while leveraging the reference image conditions to maintain consistency.

In this paper, we propose a training-based method, Lay2Story, which not only keeps the subject consistent but also enables more refined control over the subject by injecting layout conditions into the model, including its position, appearance, clothing, expression, posture, and other relevant details. Our model consists of two main branches: the global branch and the subject branch. The global branch takes noise latent as input, guided by global captions, and focuses on generating the overall image content. The subject branch takes as input the noise latent, subject mask, and latent vector of a reference image, guided by subject captions and focuses on maintaining subject consistency while generating the subject's position and detailed attributes. The Lay2Story model, built on Diffusion Transformers (DiTs), is based on the PixArt-α [2] image generation model. Inspired by MM-DiT, Lay2Story employs Masked 3D Self-Attention to enhance subject consistency through inter-frame attention guided by subject masks. Unlike StoryDiffusion, it is trained on consistent sequences; unlike Storynizor, it additionally incorporates subject information for more precise layout control. During training, we first fine-tune the base model with image data from Lay2Story-1M, then freeze the global branches and train the subject branches on a consistent frame sequence. This enables our model to simultaneously achieve consistency, semantic relevance, and aesthetic quality.

B. Examples of Lay2Story-1M

B.1. Frame Sequence Examples

As shown in Fig. 1, we provide the image data of frame sequences from the Lay2Story-1M dataset (without showing annotation information such as global captions or layout conditions), with sequence lengths ranging from 4 to 6 frames.

B.2. Examples of Lay2Story-Bench

As shown in Fig. 2, we present examples from Lay2Story-Bench, including raw frame sequence images and their corresponding annotations, which cover global captions, subject positions, and subject captions for each frame.

C. Preliminary

C.1. Latent Diffusion Models

Latent diffusion models [19] learn a denoising process to simulate the probability distribution within latent space. To reduce the computational load, the image x is transformed into a latent space feature $z_0 = E(x)$ using a Variational Autoencoder (VAE) Encoder E [6]. During the forward diffusion process, Gaussian noise is iteratively added to z_0 at timesteps t, resulting in z_t, according to the equation:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I) \tag{1}$$

where β_t represents a sequence schedule. The denoising process is defined as an iterative Markov Chain that denoises the initial Gaussian noise $z_T \in \mathcal{N}(0, I)$ into the clean latent space z_0. The denoising function in LDM is typically implemented with U-Net [20] or Transformers [26], trained by minimizing the mean squared error loss:

$$L = \mathbb{E}_{z_t,c,t,\epsilon\sim\mathcal{N}(0,I)}\left[\|\epsilon - \epsilon_\theta(x_t; c, t)\|_2^2\right] \tag{2}$$

where ϵ_θ represents the parameterized network for predicting noise, and c denotes an optional conditional input. Subsequently, the denoised latent space feature is decoded into image pixels using the VAE Decoder D.

C.2. Diffusion Transformers

In the task of consistent image generation, improvements are often made to the U-Net model [20], with common optimizations including SD1.5 [19] and SDXL [17]. In recent years, Transformer-based approaches have gradually matured in the field of text-to-image generation, with representative methods such as Stable Diffusion 3 [3] and PixArt-α [2]. These methods have demonstrated the significant advantages of Diffusion Transformers in terms of scalability, an area where U-Net falls short. The core module of PixArt-α consists of three parts: first, the linear layers that generate scale shift parameters for output normalization; second, a self-attention mechanism with latent inputs to enhance generation quality; and third, a cross-attention mechanism that takes both latent and text embeddings as inputs, using textual information as a condition to guide the generation process.

D. Training and Inference Settings

We adopt a similar approach to PixArt-α, using T5 [18] as the text encoder with a fixed token length 120. The training process consists of two stages. In the first stage, we fine-tune the global branch for the text-to-image task, training the model with the AdamW optimizer at a learning rate of 2e-5 and a weight decay of 0.03. The model runs for 5 epochs on the Lay2Story-1M dataset using 16 40GB A100 GPUs. In the second stage, we freeze the global branch and train the subject branch of Lay2Story independently, using the AdamW optimizer with a learning rate of 1e-5 and the same weight decay. This stage lasts for 10 epochs with 32 80GB A100 GPUs. During inference, we follow the configuration of previous studies, using 25 sampling steps and setting the class-free guidance coefficient to 4.5.

E. Supplementary Analyses and Experiments

E.1. Computational Complexity Analysis

Table 1 reports GPU memory usage and inference time.

Table 1. **Computational cost**. All experiments were conducted on an 80GB A100 GPU using FlashAttention at a resolution of 720p.

Frame num	Inference time (s)	Memory (MiB)
4	14.02	29731
8	17.70	33072
16	29.69	46320
32	78.33	62127

E.2. Multi-Subject Experiments

Owing to the high cost associated with data collection and training, the current experiments are limited to single-subject narratives. Nonetheless, the proposed pipeline is inherently compatible with multi-subject scenarios, as it preserves all subject bounding boxes during the Grounding DINO detection stage, followed by feature extraction, clustering, and grouping. Multi-subject handling is facilitated

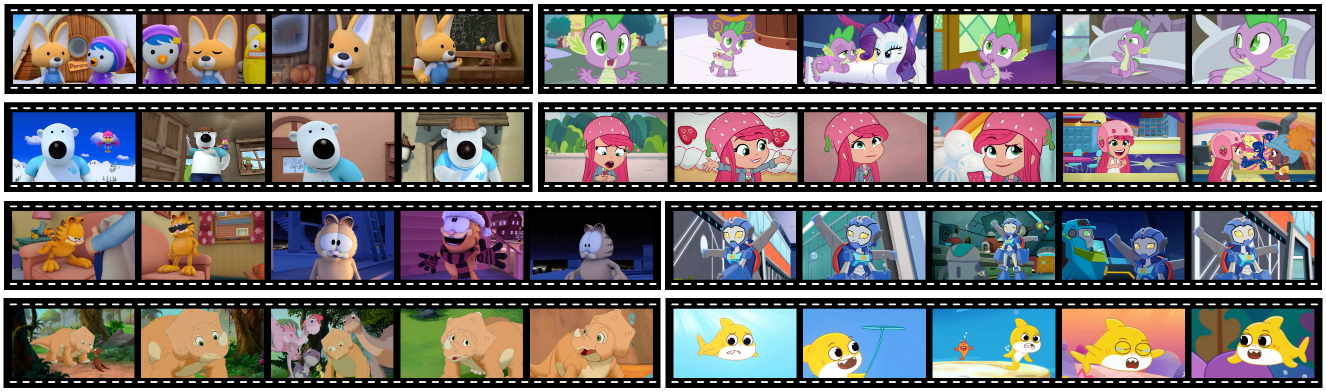
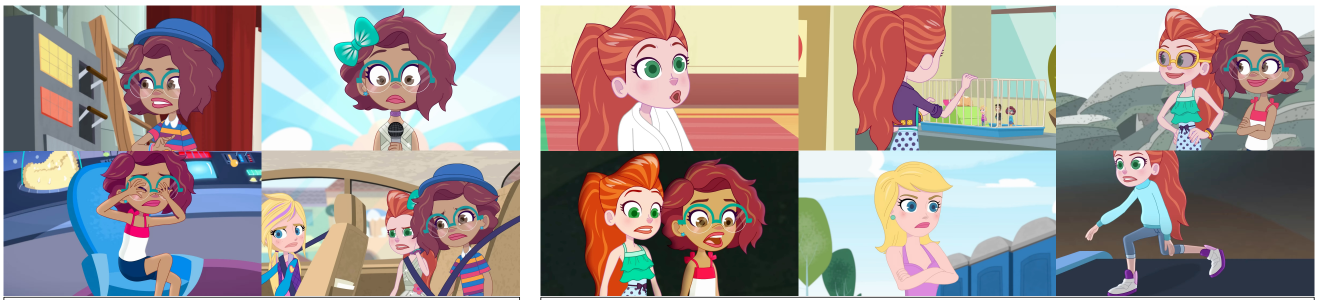


Figure 1. **Frame sequence examples.** We present renderings of several frame sequences from Lay2Story-1M.



Frame 1:
{
Global captions: A girl with red hair is standing by a ladder in the room,
Subject positions: [788, 142, 1090, 720],
Subject captions: green glasses, red curly hair, blue hat, colorful short sleeves, surprised expression,
}
Frame 2:
{
Global captions: A girl with red hair is standing on stage with a microphone against a backdrop of blue sky and white clouds and bright light,
Subject positions: [421, 22, 986, 720],
Subject captions: green glasses, green knot, strange expression,
}
...

Frame 1:
{
Global captions: A young girl whistling indoors in a gymnasium,
Subject positions: [144, 0, 764, 720],
Subject captions: red hair, green pupils, white tracksuit, surprised expression,
}
Frame 2:
{
Global captions: A young girl standing in a room hands on a railing,
Subject positions: [315, 0, 803, 720],
Subject captions: red hair, green pupils, white tracksuit,
}
Frame 3:
{
Global captions: A young girl standing on a rock beside another girl,
Subject positions: [570, 0, 871, 720],
Subject captions: red hair, yellow sunglasses, little green suspenders
}
...

Figure 2. **Examples of Lay2Story-Bench.** We present examples from the Lay2Story-Bench benchmark, including the original images and annotations, which consist of global captions, subject positions, and subject captions for each frame.

by concatenating the positional embeddings of all subjects and conditioning the model on the corresponding textual embeddings, thereby maintaining spatial layout and texture consistency. Comprehensive evaluation under multi-subject settings is left as an avenue for future exploration.

References

[1] BlackForestlabs AI. Flux. https://github.com/black-forest-labs/flux, 2024. Accessed: March 6, 2025. 1

[2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion

transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1, 2

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

[5] Biao Gong, Siteng Huang, Yutong Feng, Shiwei Zhang, Yuyuan Li, and Yu Liu. Check locate rectify: A training-free layout calibration system for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6624–6634, 2024. 1

[6] Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019. 2

[7] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15049–15058, 2021. 1

[8] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 1

[9] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6180–6189, 2024. 1

[10] Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7797–7806, 2024. 1

[11] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

[12] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 1

[13] Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. *arXiv preprint arXiv:2501.13554*, 2025. 1

[14] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 1

[15] Yuhang Ma, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *Advances in Neural Information Processing Systems*, 37:128886–128910, 2024. 1

[16] Saman Motamed, Danda Pani Paudel, and Luc Van Gool. Lego: Learning to disentangle and invert personalized concepts beyond object appearance in text-to-image diffusion models. *arXiv preprint arXiv:2311.13833*, 2023. 1

[17] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[21] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1

[22] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, et al. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36:66860–66889, 2023. 1

[23] InstantX Team. Instantx flux.1-dev ip-adapter page, 2024. 1

[24] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 1

[25] Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 1

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[27] Anton Voronov, Mikhail Khoroshikh, Artem Babenko, and Max Ryabinin. Is this loss informative? faster text-to-image customization by tracking objective dynamics. *Advances in Neural Information Processing Systems*, 36:37491–37510, 2023. 1

[28] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 1

[29] Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A unified stylized image generation model. *arXiv preprint arXiv:2309.01770*, 2023. 1

[30] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14256–14266, 2023. 1

[31] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 1

[32] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024. 1

[33] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 1

[34] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1

[35] Hui Zhang, Dexiang Hong, Tingwei Gao, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. *arXiv preprint arXiv:2412.03859*, 2024. 1

[36] Xulu Zhang, Xiao-Yong Wei, Wengyu Zhang, Jinlin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A survey on personalized content synthesis with diffusion models. *arXiv preprint arXiv:2405.05538*, 2024. 1

[37] Xulu Zhang, Wengyu Zhang, Xiaoyong Wei, Jinlin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Generative active learning for image synthesis personalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10669–10677, 2024. 1

[38] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8584–8593, 2019. 1

[39] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 1

[40] Yufan Zhou, Ruiyi Zhang, Tong Sun, and Jinhui Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*, 2023. 1

[41] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37: 110315–110340, 2025. 1