Accelerated Training-Free Character-Consistent Text-to-Image Generation Framework

Doyun Kwon
Department of Applied Artificial Intelligence
Seoul National University of Science and Technology
Seoul, Republic of Korea
vividsu1@seoultech.ac.kr

Geonoh Nam
Department of Applied Artificial Intelligence
Seoul National University of Science and Technology
Seoul, Republic of Korea
geonohn@seoultech.ac.kr

Mingyoo Song
Department of Applied Artificial Intelligence
Seoul National University of Science and Technology
Seoul, Republic of Korea
mgsong99@seoultech.ac.kr

Hanul Kim
Department of Applied Artificial Intelligence
Seoul National University of Science and Technology
Seoul, Republic of Korea
hukim@seoultech.ac.kr

***Abstract*—We tackle character-consistent text-to-image (T2I) generation in a training-free setting. Shared attention with a subject mask enforces identity consistency across prompts but introduces substantial memory and latency overhead at inference. To mitigate this cost, we augment the stable diffusion baseline with two accelerations: token pruning, which removes redundant tokens, and adaptive guidance, which skips unnecessary computation during the diffusion process. Experiments on the ConsiStory+ benchmark show that our method outperforms recent state-of-the-art approaches in character-consistent T2I generation. Notably, it attains lower inference latency than both prior state-of-the-art methods and the SDXL baseline.**

***Index Terms*—Text-to-image generation, character-consistent image generation, efficient image generation.**

I. Introduction

Recent advances in generative models [1], [2] enable high-fidelity image generation. Consequently, text-to-image (T2I) models [3]–[5] allow users to generate desired realistic images through text prompt conditions. These T2I models enable users to generate novel scenes with previously unseen combinations and generate vivid images across diverse styles [6]. However, T2I models face persistent difficulties in maintaining character consistency across different text prompt conditions [7].

The character consistency problem in T2I models refers to the failure of T2I models to maintain consistent appearances when generating a series of images from multiple text prompts with the same character. To address this, recent studies [7]–[9] propose training-free approaches that enhance character consistency by leveraging tokens from other frames during self-attention in T2I diffusion models. These methods enable image generation with consistent characters, but require extensive memory resources or complex module designs [10], both of which lead to increased inference time. Therefore, acceleration methods for T2I diffusion models need to be combined with methods that improve character consistency.

Building on recent advances in diffusion acceleration [11]–[13], we can explore how these methods can be adapted with character-consistent generation methods. Recently, training-free acceleration methods have been proposed to reduce computational costs during inference. First, token pruning method [14] reduces the number of tokens, considering that the computational cost of attention layers within diffusion models increases quadratically with the number of tokens [15]. Second, enhancing classifier-free guidance [16] (CFG) method [17] can improve the sampling efficiency of diffusion models.

In this work, we apply methods to maintain character consistency while accelerating image generation. Specifically, we apply shared attention [8], where current frame query tokens attend to key and value tokens across other frames. However, shared attention tends to homogenize backgrounds across frames. To preserve background diversity, we apply a subject mask [8] that restricts shared attention to character regions in other frames. While shared attention ensures character consistency, it increases inference time. To address this, we employ a token pruning method [14] that prunes the redundant tokens. We also apply adaptive guidance [17] that skips half of the unconditional passes in CFG. By combining these components, we achieve both character consistency and reduced inference time. We validate our approach on the ConsiStory+ benchmark [10] and compare against state-of-the-art methods [8], [10], with detailed results in Section IV.

II. Related Work

A. Character Consistency

Character consistency remains a challenging problem when generating images across different text prompt conditions in text-to-image (T2I) generation methods [3]–[5]. T2I methods

This work was supported in part by the Ministry of SMEs and Startups of Korea (MSS), under Grant No. RS-2025-02316362, and by the National Research Foundation of Korea (NRF) under Grant No. RS-2023-00221365.

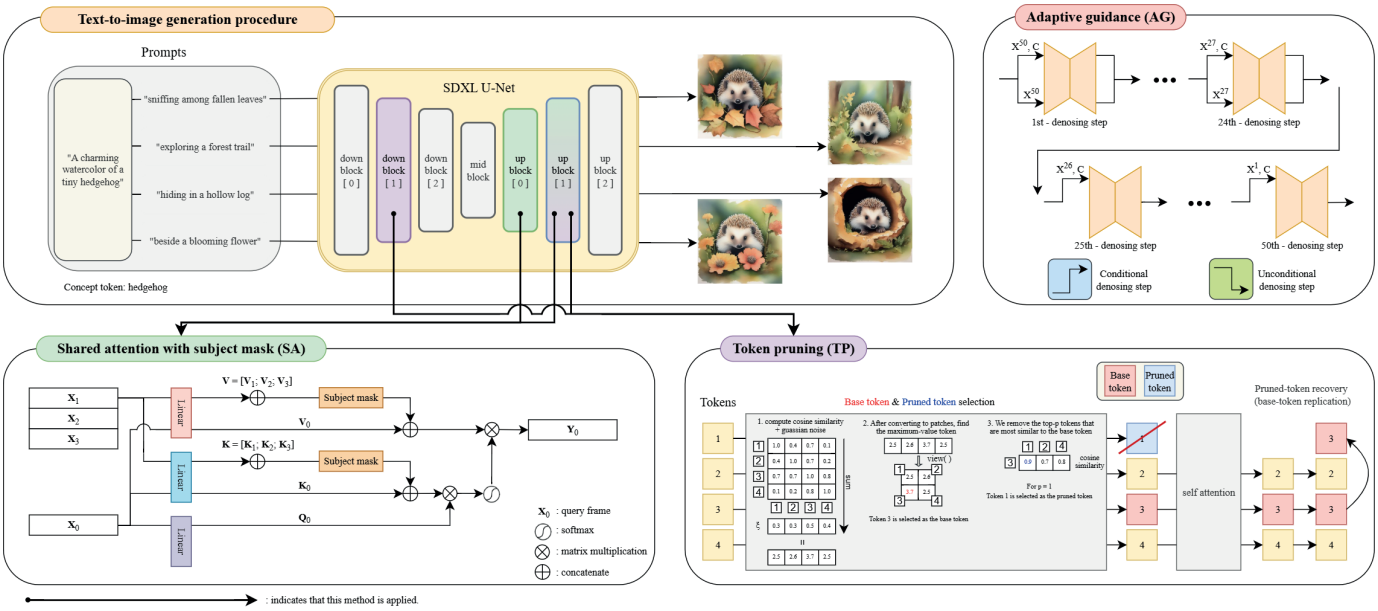

Fig. 1. An overview of our method, which consists of three main components: (1) Shared attention with subject mask to ensure character consistency; (2) Token pruning to prune redundant tokens; (3) Adaptive guidance to reduce CFG computation.

inherently generate each sample independently, making it difficult to maintain character consistency when generating images of the same character. This limitation has also led to difficulties in image personalization [18], [19] and story-telling applications [20]–[22] both of which require character consistency. Recent T2I diffusion models handle character consistency via three main approaches: (1) embedding and fine-tuning based approaches [23]–[27], which adapts models to user-specific data by directly modifying model parameters or learning subject-specific embeddings; (2) prompt-based approach [10] that enables maintaining consistent character while generating backgrounds and other elements according to each prompt when concatenating multiple prompts into one; and (3) attention-based approaches [7]–[9], which introduce self-attention consistency modules or shared attention blocks that propagate character information across frames.

B. Accelerating Image Generation

Image generation speed, along with improvements in character consistency, is one of the critical factors to consider in T2I diffusion models. However, the trade-off between performance and efficiency presents a significant challenge in diffusion models [28]. The primary causes of generation slowdowns stem from the diffusion model's intrinsic complexity and the computational demands during inference. To overcome these challenges, existing methods mainly focus on two main approaches: (1) model compression and optimization approaches [11]–[13] that reduce computational complexity by optimizing model size and architecture—such as knowledge distillation that train smaller student models to reduce the number of sampling steps, as well as neural architecture search methods that identify more efficient model designs; (2) approach of reducing computation during inference [14], [17], including adaptive guidance that selectively reduces classifier-free guidance (CFG) computation in later part of denoising timesteps and token pruning which removes redundant tokens in attention operations.

III. Proposed Method

Fig. 1 presents an overview of our pipeline for character-consistent image generation. We adopt the stable diffusion architecture [3] as the base image generator, which takes a set of input prompts and produces corresponding images via a U-Net backbone. We then incorporate shared attention and subject masking [8] to enforce character consistency while preserving semantic alignment between the generated images and the input prompts. Finally, we integrate token pruning [14] and adaptive guidance [17] into the generator to reduce inference-time cost.

A. Character-Consistent Image Generation

Shared attention. Let $\mathbf{X}_i \in \mathbb{R}^{n\times d}$ denote the image-token sequence of the i-th frame, where n is the number of tokens per frame and d is the channel dimension. We obtain $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{n\times d}$ by applying linear projections to $\mathbf{X}_i$. Aggregating all m frames yields $\mathbf{K} = [\mathbf{K}_1; \ldots; \mathbf{K}_m]$ and $\mathbf{V} = [\mathbf{V}_1; \ldots; \mathbf{V}_m] \in \mathbb{R}^{(mn)\times d}$, where $[\cdot;\cdot]$ denotes concatenation along the token dimension. The i-th-frame output $\mathbf{Y}_i$ of shared attention is given by

$$\mathbf{Y}_i = \mathrm{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \tag{1}$$

which extends self-attention by allowing the queries of frame i to attend to keys and values aggregated over all m frames. This cross-frame context couples the generation processes and

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS IN CHARACTER-CONSISTENT IMAGE GENERATION [8], [10], AND SDXL BASELINE [3]. THE BEST RESULTS ARE DEPICTED IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Methods	DreamSim ↓	CLIP-I ↑	CLIP-T ↑	Inference Time
SDXL [3]	0.3564	0.8705	<u>0.9160</u>	<u>9.8543</u>
Consistory [8]	0.2573	0.9032	**0.9236**	12.9596
1P1S [10]	<u>0.2040</u>	<u>0.9194</u>	0.8901	20.8879
Ours	**0.2039**	**0.9244**	0.8998	**9.2517**

promotes character-consistent images rather than isolated per-frame outputs. Therefore, we replace the self-attention layers in the first and second upsampling blocks of the generator with shared-attention layers.

Subject masking. While the coupled context induced by shared attention improves character consistency, it can inadvertently reduce diversity yielding similar backgrounds and layouts, and cause misalignment with the input prompts. To mitigate these side effects, we adopt subject masking [8], which restricts shared attention to the subject regions of other frames. Given a concept token indicating the subject category, we compute cross-attention weights between this token and the image tokens of frame i. Because the concept token encodes the subject, these weights are higher over subject regions. We average the weights to form a score map and binarize it with thresholding [29] to obtain the subject mask $\mathbf{M}_i$. We also perform mask dropout, randomly zeroing entries with rate α, to diversify the object-region signal during training. We then apply $\mathbf{M}_i$ to shared attention as

$$\mathbf{Y}_i = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}^\top}{\sqrt{d}} + \log \mathbf{M}_i\right)\mathbf{V} \tag{2}$$

B. Accelerated Character-Consistent Image Generation

Token pruning. Token pruning accelerates inference by reducing the number of tokens processed. For the i-th frame's tokens $\mathbf{X}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in}\}$, we compute each token's total similarity to the rest. Specifically, for the k-th token we define

$$s_{ik} = \sum_{j=1}^{n} \frac{\mathbf{x}_{ik}^\top \mathbf{x}_{ij}}{\|\mathbf{x}_{ik}\| \, \|\mathbf{x}_{ij}\|} + \xi_{ik}, \qquad \xi_{ik} \sim \mathcal{N}(0, \sigma^2). \tag{3}$$

Here, ξ_{ik} is a small Gaussian perturbation that injects stochasticity across time steps, preventing the same tokens from being repeatedly selected as bases or pruned. Following [14], we spatially partition the image tokens $\mathbf{X}_i$ into 2×2 groups. For each group, we select a single base token that maximizes the total similarity. We then prune the top-p tokens most similar to the base tokens, where p equals the total number of tokens multiplied by the pruning ratio. Attention is applied only to the remaining tokens. After attention, we restore the original length by filling each pruned position with the output of its most similar base token. We observed that additional pruning in the fully connected layer improves speed but degrades image quality. Therefore, pruning is performed only for the attention layer.

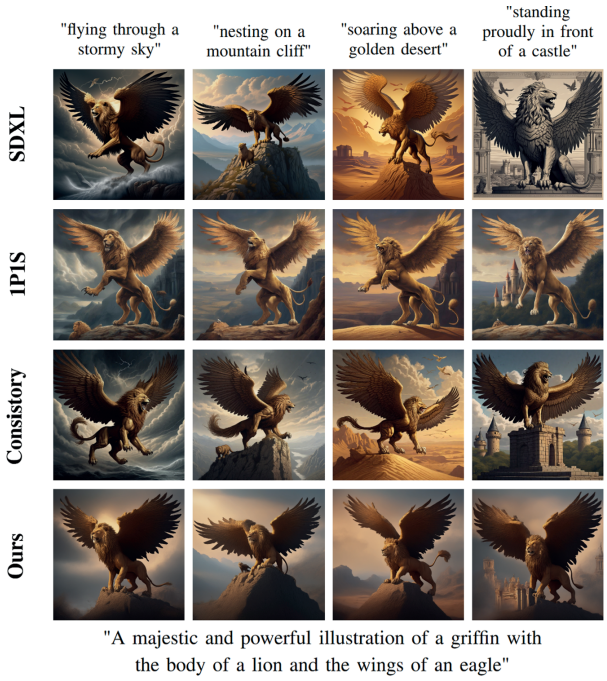

Fig. 2. Qualitative comparison with state-of-the-art methods in character-consistent image generation [8], [10] and SDXL baseline [3].

Adaptive guidance. Given input noise $\mathbf{X}^T \sim \mathcal{N}(0, \mathbf{I})$ and a text condition $\mathbf{C}$, the diffusion model progressively denoises $\mathbf{X}^T$ to produce the target image $\mathbf{X}^0$. Let $\mathbf{X}^t$ denote the image tokens at timestep $t \in \{T, \ldots, 0\}$. Under classifier-free guidance (CFG) [16], we combine the conditioned $\epsilon_\theta(\mathbf{X}^t, \mathbf{C})$ and unconditioned $\epsilon_\theta(\mathbf{X}^t, \varnothing)$ predictions from the diffusion model ϵ_θ where θ is pre-trained parameters and $\varnothing$ indicates no conditioning. CFG improves text alignment and generation stability. However, it doubles computation burden by evaluating both branches at every timestep.

As reported in [17], the benefits of CFG diminish as sampling progresses. We therefore adopt adaptive guidance [17] that skips the unconditional branch in later timesteps:

$$\epsilon_{\text{cfg}}(\mathbf{X}^t, \mathbf{C}, \omega) = \begin{cases} \epsilon_\theta(\mathbf{X}^t, \varnothing) + \omega\Delta\epsilon & \text{if } t \geq \tau \\ \epsilon_\theta(\mathbf{X}^t, \mathbf{C}) & \text{if } t < \tau \end{cases} \tag{4}$$

where $\Delta\epsilon = \epsilon_\theta(\mathbf{X}^t, \mathbf{C}) - \epsilon_\theta(\mathbf{X}^t, \varnothing)$, ω balances the guidance scale, and τ is the timestep cutoff. This reduces inference cost while preserving strong guidance when it matters most.

IV. EXPERIMENTS

In this section, we perform experiments to validate our method. We first describe the experimental setup. We then compare with state-of-the-art methods. Finally, we provide an ablation study of our method.

TABLE II
ABLATION STUDY ON THE THREE MAIN COMPONENTS. THE BEST RESULTS ARE DEPICTED IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED. THE FINAL SETTINGS ARE HIGHLIGHTED IN LIGHTGRAY.

Components			DreamSim ↓	CLIP-I ↑	CLIP-T ↑	Inference Time
Shared attention with subject mask	Token pruning	Adaptive guidance				
✗	✗	✗	0.3564	0.8705	<u>0.9160</u>	9.8543
✓	✗	✗	0.2141	0.9182	**0.9169**	12.1963
✓	✓	✗	**0.2008**	**0.9265**	0.9076	11.9568
✓	✗	✓	0.2132	0.9183	0.9124	<u>9.6894</u>
✓	✓	✓	<u>0.2039</u>	<u>0.9244</u>	0.8998	**9.2517**

A. Experimental Setup

Dataset. To evaluate prompt-image alignment, character consistency, and inference time, we utilize the ConsiStory+ benchmark [10]. ConsiStory+ contains 8 superclasses, including humans, animals, fantasy, inanimate objects, fairy tales, nature, technology, and food. Each superclass contains multiple concept tokens, where each concept token is provided with a subject description, a style description, and 5 to 10 setting descriptions. Prompts are formed by combining a fixed pairing of a subject description that defines the character and a style description that specifies its visual representation with multiple setting descriptions to create various prompts. Using the images generated from these prompts, we measure character consistency. ConsiStory+ consists of 192 concept tokens and contains a total of 1,100 prompts.

Evaluation metrics. Our evaluation protocols follow the metrics used in [10]. To evaluate character identity consistency, we measure DreamSim [30] and CLIP-I [31]. Both DreamSim and CLIP-I utilize CarveKit [32] to remove image backgrounds and replace them with random noise. The CLIP-T metric assesses prompt alignment by measuring the CLIPScore [31] between the corresponding prompt and generated image. In addition, we report inference time, which is measured in second, to assess the efficiency of image generation. The inference time is obtained by dividing the total duration of the ConsiStory+ benchmark by the total number of generated images.

Implementation details. We adopt the experimental setup of prior work [8]. Specifically, we generate 1024×1024 resolution images using pretrained SDXL [3]. During inference, we set the parameters to $\tau = 25$, $T = 50$, and $\omega = 5.0$ in (4). For the dropout applied to the subject mask, α is set to 0.5. All experiments are performed on a NVIDIA A6000 GPU.

B. Main Results

Table I presents a quantitative comparison of our method against other state-of-the-art methods. The compared methods include Consistory [33] and 1P1S [10]. Since these methods are based on SDXL, we also compare with the quantitative results of SDXL. Our method shows the best performance in DreamSim and CLIP-I, achieving scores of 0.2039 and 0.9244, respectively. Although the CLIP-T score of 0.8998 falls below Consistory and SDXL, it shows improvement over 1P1S. In terms of generation speed, our method achieves an inference time of 9.2517 seconds, running 55.7% faster than 1P1S.

Fig. 3. Qualitative results across combinations of the main components. SA and SM denote shared attention and subject mask, respectively.

Fig. 2 shows generated images using our method and state-of-the-art methods. We evaluate against Consistory [33] and 1P1S [10], along with SDXL as the baseline model. The results show that while other comparison methods generate griffins with incomplete legs or bodies, our method creates complete and consistent subjects. Fig. 3 presents a comparison of how our methods affect the quality of the generated images. In the first row, SDXL generates puppies with different appearances across frames. The second row shows the effect of applying shared attention with a subject mask. This visually confirms that improves character consistency. Also, the third row shows that adding adaptive guidance and token pruning makes very little visual difference. This result shows that adaptive guidance and token pruning, which accelerate image generation, do not substantially compromise output quality.

C. Ablation study

Table II summarizes ablation study to example the contribution of building components: shared attention with subject mask, token pruning, and adaptive guidance.

Shared attention with subject mask. Compared to the first row, the second improves DreamSim, CLIP-I, and CLIP-

T by 0.1423, 0.0477, and 0.0009, respectively, indicating better subject consistency and prompt alignment. These gains increase inference time by 23.8% due to the added cost of shared attention.

Token pruning. The third row shows the impact of token pruning. Compared to the second row, token pruning made generation 0.2395 seconds faster. In addition, token pruning can be combined with adaptive guidance to achieve even more substantial speed improvements.

Adaptive guidance. The fourth row demonstrates the impact of adaptive guidance: relative to the second row, generation is faster by 2.5069 seconds. Our final configuration (last row) applies all three components; versus the first row, DreamSim improves by 0.1525, CLIP-I by 0.0539, while CLIP-T decreases by 0.0162. Moreover, relative to the second row, the inference-time penalty of shared attention is mitigated by 24.1%, from 12.1963 seconds to 9.2517 seconds.

V. Conclusion

In this paper, we applied shared attention with subject mask, token pruning, and adaptive guidance components to enhance both character consistency and image generation speed. Experiments showed that our method achieves comparable performance with faster generation speed compared to state-of-the-art methods. Notably, our method achieved faster inference time than the SDXL baseline. Furthermore, ablation studies further confirmed the individual contributions of our components. These results showed that the generation speed degradation caused by shared attention for character consistency can be overcome by combining acceleration components with marginal performance degradation.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, vol. 27, 2014.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[3] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *ICLR*, 2024.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[5] J. Chen, S. Xue, Y. Zhao, J. Yu, S. Paul, J. Chen, H. Cai, S. Han, and E. Xie, "Sana-sprint: One-step diffusion with continuous-time consistency distillation," *arXiv preprint arXiv:2503.09641*, 2025.

[6] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *ICLR*, 2023.

[7] Y. Zhou, D. Zhou, M.-M. Cheng, J. Feng, and Q. Hou, "Storydiffusion: Consistent self-attention for long-range image and video generation," in *NeurIPS*, 2024.

[8] Y. Tewel, O. Kaduri, R. Gal, Y. Kasten, L. Wolf, G. Chechik, and Y. Atzmon, "Training-free consistent text-to-image generation," *TOG*, vol. 43, no. 4, 2024.

[9] J. Singh, J. K. Chen, J. K. Kohler, and M. F. Cohen, "Storybooth: Training-free multi-subject consistency for improved visual storytelling," in *ICLR*, 2025.

[10] T. Liu, K. Wang, S. Li, J. van de Weijer, F. S. Khan, S. Yang, Y. Wang, J. Yang, and M.-M. Cheng, "One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt," in *ICLR*, 2025.

[11] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *ICLR*, 2022.

[12] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *NeurIPS*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 26 565–26 577.

[13] C. Chadebec, O. Tasar, E. Benaroche, and B. Aubin, "Flash diffusion: Accelerating any conditional diffusion model for few steps image generation," in *AAAI*, vol. 39, no. 15, 2025, pp. 15 686–15 695.

[14] E. Zhang, J. Tang, X. Ning, and L. Zhang, "Training-free and hardware-friendly acceleration for diffusion models via similarity-based token pruning," in *AAAI*, vol. 39, no. 9, 2025, pp. 9878–9886.

[15] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," in *ICLR*, 2024.

[16] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS*, 2021.

[17] A. Castillo, J. Kohler, J. C. Pérez, J. P. Pérez, A. Pumarola, B. Ghanem, P. Arbeláez, and A. Thabet, "Adaptive guidance: Training-free acceleration of conditional diffusion models," in *AAAI*, vol. 39, no. 2, 2025, pp. 1962–1970.

[18] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023.

[19] J. Huang, J. H. Liew, H. Yan, Y. Yin, Y. Zhao, H. Shi, and Y. Wei, "Classdiffusion: More aligned personalization tuning with explicit class guidance," in *ICLR*, 2025.

[20] X. Shen and M. Elhoseiny, "Storygpt-v: Large language models as consistent story visualizers," in *CVPR*, 2025, pp. 13 273–13 283.

[21] S. Zheng and Y. Fu, "Contextualstory: Consistent visual storytelling with spatially-enhanced and storyline context," in *AAAI*, vol. 39, no. 10, 2025, pp. 10 617–10 625.

[22] F. Shen, H. Ye, S. Liu, J. Zhang, C. Wang, X. Han, and Y. Wei, "Boosting consistency in story visualization with rich-contextual conditional diffusion models," in *AAAI*, vol. 39, no. 7, 2025, pp. 6785–6794.

[23] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2023.

[24] Y. Gong, Y. Pang, X. Cun, M. Xia, Y. He, H. Chen, L. Wang, Y. Zhang, X. Wang, Y. Shan *et al.*, "Talecrafter: Interactive story visualization with multiple characters," *arXiv preprint arXiv:2305.18247*, 2023.

[25] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[26] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo, "Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation," in *ICCV*, 2023, pp. 15 943–15 953.

[27] O. Avrahami, A. Hertz, Y. Vinker, M. Arar, S. Fruchter, O. Fried, D. Cohen-Or, and D. Lischinski, "The chosen one: Consistent characters in text-to-image diffusion models," in *SIGGRAPH*, 2024, pp. 1–12.

[28] Z. Wang, Y. Jiang, H. Zheng, P. Wang, P. He, Z. Wang, W. Chen, and M. Zhou, "Patch diffusion: Faster and more data-efficient training of diffusion models," in *NeurIPS*, 2023.

[29] N. Otsu, "A threshold selection method from gray-level histograms," *TSMC*, vol. 9, no. 1, pp. 62–66, 1979.

[30] S. Fu, N. Y. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "Dreamsim: Learning new dimensions of human visual similarity using synthetic data," in *NeurIPS*, 2023.

[31] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *EMNLP*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., 2021, pp. 7514–7528.

[32] N. Selin, "Carvekit: Automated high-quality background removal framework," 2023. [Online]. Available: https://github.com/OPHoperHPO/image-background-remove-tool

[33] L. Li, H. Li, X. Zheng, J. Wu, X. Xiao, R. Wang, M. Zheng, X. Pan, F. Chao, and R. Ji, "Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration," in *ICCV*, 2023, pp. 7105–7114.