Personalized Text-to-Image Generation with Attribute Disentanglement and Feature Embedding

1st Xiangyu Hou
Xi'an Jiaotong University
Xi'an, China
houxyben@stu.xjtu.edu.cn

***Abstract*—Text-to-image generation has made remarkable progress in recent years, yet achieving personalized and consistent generation remains a significant challenge. In this paper, we propose a novel framework that combines ProSpect-inspired multi-stage learning and attribute disentanglement with advanced feature extraction and embedding techniques to address these challenges. Our method leverages a diffusion-based architecture to generate high-fidelity images conditioned on both textual prompts and user-provided reference images. By adopting a multi-stage approach, the model progressively refines the image, ensuring alignment with global structures, specific attributes, and fine-grained details. Attribute disentanglement enables precise control over visual characteristics such as style, color, and structure, while feature extraction and embedding mechanisms ensure accurate representation of user-specific concepts. Our approach requires only a single reference image, making it highly practical and scalable. Extensive experiments demonstrate that our method outperforms existing approaches in terms of image quality, personalization accuracy, and cross-generation consistency. Additionally, our framework offers strong editability, allowing users to modify specific attributes without compromising overall quality. This work advances the state-of-the-art in text-to-image generation, providing a robust and flexible solution for personalized and consistent image creation in creative applications.**

***Index Terms*—Text-to-Image, Diffusion,ProSpect**

I. Introduction

Text-to-image generation has emerged as a transformative technology in artificial intelligence [1], [23], enabling the creation of visual content from natural language descriptions. Recent advancements in diffusion models have significantly enhanced the quality, diversity, and realism of generated images, making them a cornerstone of modern generative AI. Unlike traditional generative adversarial networks (GANs) [2], diffusion models excel at capturing fine-grained details and producing high-fidelity images through an iterative denoising process. Technologies such as GLIDE [3], DALL-E 2 [4] , Imagen [5], Stable Diffusion (SD) [6], eDiff-I [7] and RAPHAEL [8] have achieved impressive success in Text-to-image generation, Users can write text prompt to generate images with these powerful text-to-image diffusion models.

However, a critical challenge in text-to-image generation lies in achieving personalization and consistency [9]–[11]. Personalized generation aims to create images that align with specific user preferences, styles, or concepts, while consistency ensures that generated images maintain coherent attributes across multiple outputs, such as preserving the main features of a subject. These requirements are particularly important in applications like scenarios such as book illustration, game development, asset design, and advertising.

This paper addresses these challenges by proposing a novel framework built on diffusion models, specifically tailored for personalized and consistent text-to-image generation. Our approach leverages the multi-stage learning and attribute disentanglement techniques inspired by the ProSpect [12] method, combined with feature extraction and embedding mechanisms, to achieve fine-grained control over the generation process. By incorporating user-defined concepts and constraints into the diffusion pipeline, our model can generate images that not only adhere to textual prompts but also reflect unique user preferences and maintain consistency across multiple generations. Furthermore, our method requires only a single reference image, making it highly practical and accessible for real-world applications.

The contributions of this work are threefold: (1) We propose a diffusion-based architecture optimized for personalized and consistent image generation, which integrates ProSpect-inspired multi-stage learning and attribute disentanglement to achieve fine-grained control over the generation process. This enables the model to progressively refine images while maintaining high fidelity and user-specific attributes. (2) We introduce a novel conditioning mechanism that combines feature extraction and embedding techniques, allowing the model to capture and replicate the unique characteristics of a single reference image, ensuring both personalization and cross-generation consistency. This significantly reduces the reliance on multiple reference images or extensive fine-tuning. (3) We conduct extensive experimental validation, demonstrating the superiority of our approach in terms of image quality, personalization accuracy, and consistency across diverse prompts. Additionally, our method offers exceptional editability, enabling users to easily modify specific attributes of generated images without compromising overall quality.

II. Related Work

A. Text-to-Image Generation

Text-to-image generation has seen significant advancements in recent years, driven by the development of deep learning models. Early approaches relied on Generative Adversarial

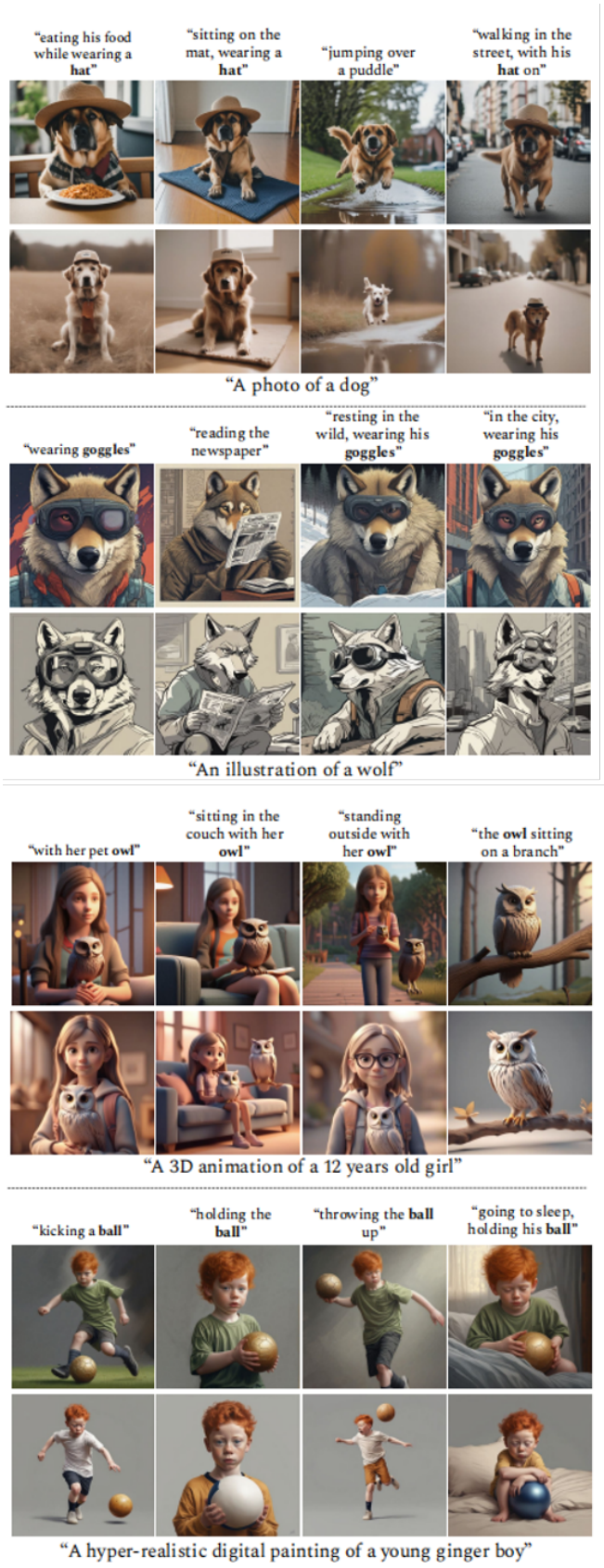

Fig. 1. Consistent character generation :Given a text prompt describing a character, our method distills a representation that enables consistent depiction of the same character in novel contexts.

Networks (GANs) [2]such as StackGAN [13] and AttnGAN [14], which generated images from textual descriptions by progressively refining low-resolution images. However, GAN-based methods often struggled with generating high-resolution images and maintaining fine-grained details, as well as suffering from mode collapse and training instability. The advent of diffusion models, such as DALL·E 2 [4], Imagen [5], and Stable Diffusion [6], has revolutionized the field by leveraging iterative denoising processes to produce high-fidelity, diverse, and photorealistic images. Stable Diffusion [6], in particular, has gained widespread attention for its open-source nature and efficient latent space modeling, which enables high-quality image generation with reduced computational costs. These models excel at capturing complex semantic relationships between text and images, enabling the generation of visually coherent and contextually relevant outputs. Despite their success, challenges remain in achieving personalized and consistent generation, particularly when adapting to user-specific concepts or maintaining consistency across multiple generations. Stable Diffusion [6] has laid a strong foundation for addressing these challenges, but further advancements are needed to fully realize the potential of text-to-image generation in practical applications.

B. Text-to-Image Personalization

Personalized text-to-image generation aims to create images that align with user-specific preferences, styles, or concepts. Early methods, such as textual inversion [9], introduced techniques to embed user-defined concepts into the latent space of pre-trained models, allowing for the generation of images that reflect specific attributes. DreamBooth [11] further advanced this field by fine-tuning diffusion models on a small set of user-provided images, enabling the generation of personalized content with high fidelity. However, these methods often require multiple reference images or extensive fine-tuning, limiting their scalability and practicality. Recent approaches have sought to address these limitations through more efficient and flexible techniques. For instance, Custom Diffusion and Concept Sliders explore lightweight fine-tuning and modular adjustments to improve personalization efficiency. Additionally, IP-Adapter [15] leverages adapter-based architectures to enable fast adaptation to new concepts without retraining the entire model. ELITE [16] further enhances personalization by disentangling visual attributes and enabling fine-grained control over specific features, such as style or texture. Despite these advancements, achieving consistent and editable personalized generation with minimal user input remains an open challenge. These methods have significantly improved the efficiency and flexibility of personalization, but further research is needed to fully integrate user-specific concepts into the generation process while maintaining high-quality outputs and cross-generation consistency.

C. Image Editing

Image editing in the context of text-to-image generation focuses on modifying generated or existing images while

preserving their core attributes. Early techniques relied on GAN-based inpainting and style transfer methods, which allowed for localized edits but often lacked precision and semantic understanding. The introduction of diffusion models has enabled more sophisticated editing capabilities, as seen in Prompt-to-Prompt [17] and DiffEdit [**?**], which leverage cross-attention mechanisms to align textual prompts with image regions for targeted modifications. Additionally, methods like InstructPix2Pix [19] and Text2LIVE [20] have explored instruction-based editing, enabling users to guide edits through natural language instructions. However, these approaches often struggle with maintaining consistency across edits or require complex user inputs. Our work builds on these foundations by integrating attribute disentanglement and multi-stage learning, enabling fine-grained and consistent image editing with minimal user effort.

III. PRELIMINARIES

A. Text-to-image diffusion models

Before introducing our method, we first give a brief review of ordinary text-to-image diffusion models [21]. To start with, Gaussian diffusion models assume a forward Markov process that gradually adds noise to normal image x_0 :

$$x_t = \sqrt{\bar{\alpha}_0} x_t + \sqrt{\bar{\alpha}_t}\epsilon \tag{1}$$

where $t \in [0, T]$, $\epsilon \sim N(0, I)$ and $\bar{\alpha}_t$ are a set of constants. Meanwhile, a denoising network ϵ_θ, usually a U-Net [11], is trained to reverse the forward process by estimating the noise given a corrupted image:

$$\mathcal{L}(\theta) = \mathbb{E}_{t\in[0,T],x_0,\epsilon_t}[\|\epsilon_t - \epsilon_\theta(x_t, t)\|^2] \tag{2}$$

Once trained, we can sample images x_0 from a Gaussian noise x_t by gradually removing the noise step by step with ϵ_θ. The training process of the diffusion model can be seen in Fig.2

As a widely used diffusion model, Stable Diffusion (SD) [6] conducts the diffusion process in the latent space. Given an image and a text prompt, SD encodes them into latent code z and condition embedding c_t utilizing VAE [25] and CLIP [24] text encoder, respectively. In zero-shot image personalization architectures like IP-adapter, images can also be considered a condition of the diffusion model. Specifically, a subject image is encoded to image embeddings by an image encoder and then projected into the original condition space of the diffusion model denoted as c_i. For a timestep t which is uniformly sampled from a fixed range, the model θ predicts the noise ϵ_θ and is optimized through the objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t\in[0,T],z,c,\epsilon}[\|\epsilon_t - \epsilon_\theta(z_t|c_t, c_i, t)\|^2] \tag{3}$$

B. Cross-attention

In SD , both c_i and c_t are integrated into the U-Net backbone through cross-attention layers:

$$\begin{aligned} Attn(Q, K_i, K_t, V_i, V_t) &= \gamma z_{img} + z_{txt} \\ = \gamma Softmax(\frac{QK_i}{\sqrt{d}})V_i &+ Softmax(\frac{QK_t}{\sqrt{d}})V_t \end{aligned} \tag{4}$$

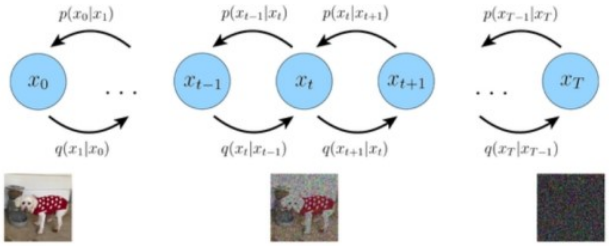

Fig. 2. The forward process(add noise) and reverse process(denoise) of diffusion model.

where $Q = zW_q$, $K_i = c_iW_k^i$, $K_t = c_tW_k^t$,$Vi = c_iW_v^i$,$Vt = c_tW_v^t$, and W_q, W_k, $W_v$are corresponding projection weight matrices. And d represents the dimensionality of the key vectors Note that the key and value matrix of c_i and c_t are independent of each other to decouple conditions of different modalities. Previous studieshave found that attention maps $A = Softmax(QK/\sqrt{d})$can reflect the attribution relation between generated images and conditions, which means that they determine the effect of condition controls

IV. METHOD

Our proposed method for achieving consistent and personalized text-to-image generation builds on the ProSpect [12] framework, incorporating its multi-stage learning and attribute disentanglement techniques, while enhancing it with feature extraction and embedding mechanisms. The approach is designed to ensure fine-grained control over the generation process, high fidelity, and consistency across multiple outputs. Below, we detail the key components of our method.

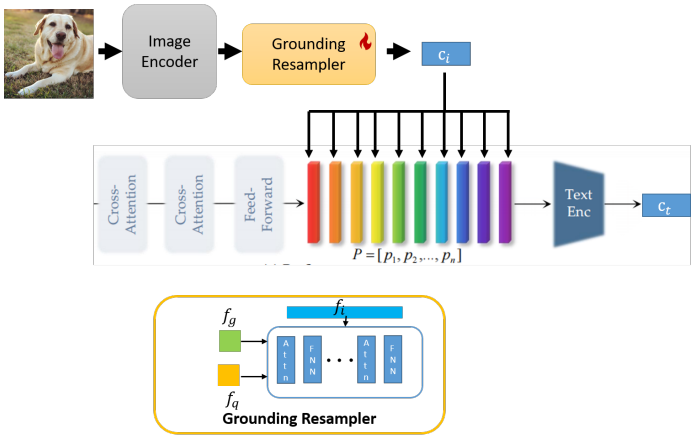

Fig. 3. The overall pipeline of our Method ,ProSpect pipeline learns a set of token embedding P=$[p_1, P_2, ..., p_n]$

The grounding resampler [22] adeptly assimilates visual information, correlating it with specific entities.Then a targeted cross-attention mechanism facilitates precise interactions between the image condition and the diffusion latent within the attention layers. constraints.

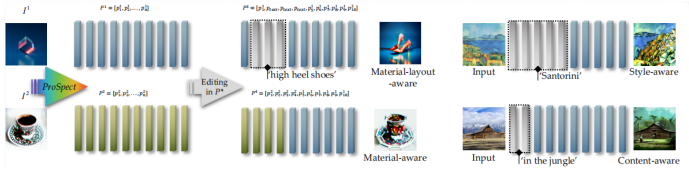

Fig. 4. Illustrations of various attribute-aware image generation tasks.

A. ProSpect

The initial generation stages of the diffusion model tend to generate overall layout and color, the middle stages tend to generate structured appearances, and the final stages tend to generate detailed textures [12]. Diffusion and denoising within Latent Diffusion Model (LDM) typically take 1000 steps,and the text conditions the model step by step. Previously, the process of the textual conditions acting on the diffusion model is regarded as a whole. In this work, we treat them as different procedures. Specifically, we divide the 1000 steps of conditioning into ten stages on average. Each stage corresponds to a unique textual condition. The collection of textual conditions reside in the CLIP [24] text-image space, their sizes are set to $n \times 1 \times 768$ (n= 10 denotes the number of the stages). This way of division is designed to keep a balance between efficiency and quality.

We refer to the expanded space as Prompt Spectrum Space, denoted as $\mathcal{P}^*$.$\mathcal{P}^*$is defined as:

$$\mathcal{P}^* = \{p_1, p_2, ..., p_n\} \tag{5}$$

where p_i represents the token embedding corresponding to the conditional prompt of the ith stage of the generation process.

$ProSpect$ is a method that maps an image to a collection of corresponding textual token embeddings.the ProSpect loss of LDM in $\mathcal{P}^*$ space is formulated as:

$$\mathcal{L}_{PS} = \mathbb{E}_{z,t,p}[\|\epsilon - \epsilon_\theta(z_t, p_i, t)\|_2^2] \tag{6}$$

where p_i =$P(t)$ is a learnable vector represents the token embedding of stage i, and P= $[p_1, p_2, ..., p_n]$ is the set of textual token embeddings in $\mathcal{P}^*$ space.

As shown in Fig.3, the token embedding is initialized to a frozen 1 × 768 text embedding with a user input text (e.g., “cup”) via the CLIP [24] text encoder. It is then fed into a randomly initialized hypernetwork and finally creates a n× 1 × 768 embedding P = $[p_1, p_2, ...p_n]$. Only the hypernetwork is trainable and the final p_i is obtained by optimizing based on Eqn. (6). The training process typically requires 1000-3000 iterations. Dropout is applied to prevent overfitting and the rate is set to 0.1.

Attribute control during inference is achieved by replacing the p_i representing different attributes with editing texts. For instance, in Fig. 4, content personalization involves maintaining the content related p_3 $p_1 0$ of image barn as “*in the jungle” and replacing p_1 p_2 with “in the jungle” (without “*”).

B. Feature Extraction

In image generation tasks, extracting key details from reference images and integrating them with textual descriptions and layout information is a significant challenge. To address this, we propose the Grounding Resampler [22], an innovative feature extraction and fusion mechanism. It uses a set of learnable query vectors to interact with image features, focusing on specific regions to extract task-relevant details. Specifically, with an image embedding f_i and a learnable query f_q, the resampler comprises several attention layers.

$$RSAttn = Softmax(\frac{Q(f_q)K([f_i, f_q])}{\sqrt{d}})V([f_i, f_q]) \tag{7}$$

where $[f_i,\ f_q]$ denotes the concatenation of the image embedding f_i and the learnable query f_q. The architecture incorporates fully connected feedforward networks (FFNs), analogous to those utilized in standard vision transformers [26].

For example, when generating ”a dog wearing pink glasses,” the query vectors concentrate on the dog’s face and glasses, capturing details like color, shape, and texture. The Resampler incorporates semantic information from text descriptions by encoding the text into embeddings, guiding the model to focus on features that match the textual description. It also uses positional information, such as bounding box coordinates, to ensure the extracted features correspond to specific areas in the image. The extracted features are combined into a single condition input for the diffusion model. During training, the Resampler randomly replaces grounding tokens with generic learnable queries to prevent over-reliance on explicit layout information, enhancing generalization. This mechanism preserves fine details, such as the color of glasses or the texture of fur, and allows precise control over each subject’s position and appearance, avoiding issues like subject overlap or incorrect placement. In summary, the Grounding Resampler bridges image features, text descriptions, and layout information, enabling precise and flexible control over multi-subject image generation.

V. EXPERIMENTS

We compare our approach with the most relevant personalization techniques. In each experiment, each technique is used to extract a character from a single image generated by SDXL [27] from the input prompt p. The same prompt p was also provided as input to our methods. Text Inversion (TI) [9] uses the same concept of several image-optimized text tokens, which we convert to support SDXL [27] by learning two text tokens (one for each text encoder), as we did in our approach. In addition, we used LoRA DreamBooth [29], which we found to be less prone to overfitting than the standard DB model. In addition, we compared all available image encoder techniques that encode individual images into the text space of the diffusion model for the next generation in the new environment: the BLIP-Diffusion [28], ELITE [16], and IP-adapter. for all baselines, we used the same prompt p to generate individual images and used it to pass the optimization

(TI and LoRA DB) or encoding (ELITE, BLIP-diffusion and IP-adapter) to extract the identity. In Fig. 5, we qualitatively compare our approach to the baseline described above.TI, BLIP-diffusion [28] and IP-adapter [15] are able to follow the specified prompts, but they are unable to generate consistent characters. LoRA DB succeeds in consistent generation, but it does not always respond to the prompts. In addition, the generated characters are generated in the same fixed poses. elite struggles with fast-following, and the generated characters tend to morph. In contrast, our approach is able to follow prompts and maintain consistency while generating engaging characters in different poses and viewpoints. To automatically

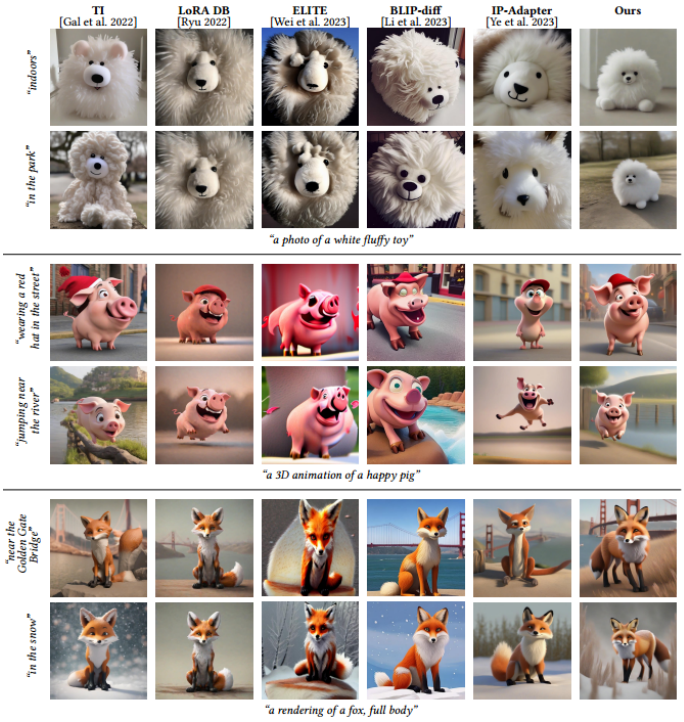

Fig. 5. We compare our method against several baselines: TI [9], BLIP-diffusion [28] and IP-adapter [15] are able to follow the target prompts, but do not preserve a consistent identity. LoRA DB [29] is able to maintain consistency, but it does not always follow the prompt. Furthermore, the character is generated in the same fixed pose. ELITE [16] struggles with prompt following and also tends to generate deformed characters. On the other hand, our method is able to follow the prompt and maintain consistent identities, while generating the characters in different poses and viewing angles.

and quantitatively evaluate our method and baseline, we instructed ChatGPT [30] to generate prompts for different types of characters (e.g., animals, creatures, objects, etc.) in different styles (e.g., stickers, animations, realistic images, etc.). Each of these prompts is then used to extract consistent characters through our method and each baseline. Next, we generate these characters in a predefined set of novel contexts. For a visual comparison, see the Supplementary Information. We used two standard evaluation metrics: prompt similarity and identity consistency, which are commonly used in the personalization literature. prompt similarity measures the correspondence between the generated image and the input text prompt. We use the standard CLIP [24] similarity, which is the normalized cosine similarity between the CLIP image embedding of the generated image and the CLIP text embedding of the source prompt. To measure identity consistency, we computed the two-by-two similarity between the CLIP image embeddings of the generated images for the same concepts in different contexts (i.e., when different textual prompts with the same characters are used). As shown in Figure 6 (left), there is an inherent trade-off between prompt similarity and identity consistency: LoRA DB and ELITE exhibit a high degree of identity consistency at the expense of prompt similarity. ti and bp -diffusion have a high degree of prompt similarity but a low degree of consistency. Our approach achieves better identity consistency than IP-adapter, which is important from the user's point of view and is supported by our user study.

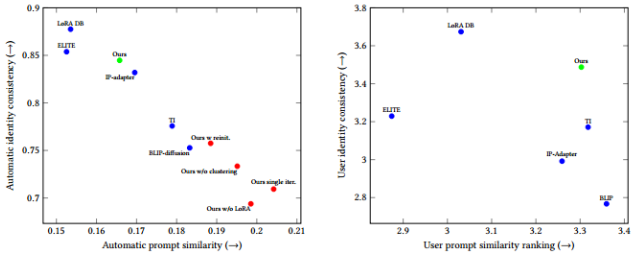

Fig. 6. Quantitative Comparison and User Study.

VI. LIMITATIONS AND CONCLUSIONS

We found the following limitations of our method:

(a) Inconsistent auxiliary roles/elements - While our approach is able to find consistent identities for the roles described by the input prompts, the identities of other roles associated with the input roles (e.g., their pets) may not be consistent. Additionally, our framework does not support finding multiple concept

(b) dummy attributes at the same time - we have found that in some cases, our approach binds additional attributes that are not present in the input text prompt to the final identity of the character. One way to mitigate this problem is to allow the user to choose the most cohesive clustering based on their preferences, rather than automatically selecting it.

(c) Simplified Characters - We found that our approach tends to generate simplified scenes (single and mostly centered objects), which may be caused by the "averaging" effect in the identity extraction phase.

(d) Use ProSpect as our basic framework can achieve attribute disentanglement, but the attribute transfer between images with large domain gap may not be visually aesthetic.

In summary, in this paper we provide a first fully automated solution to the problem of consistent character generation. We hope that our work will pave the way for future advances, as we believe that this consistent character generation technique may have a disruptive impact on education, storytelling, entertainment, fashion, branding, advertising, and many other fields.

REFERENCES

[1] J. Betker et al., "Improving image generation with better captions," Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf, vol. 2, no. 3, p. 8, 2023.

[2] I. Goodfellow et al., "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[3] A. Nichol et al., "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," arXiv preprint arXiv:2112.10741, 2021.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, vol. 1, no. 2, p. 3, 2022.

[5] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," Advances in neural information processing systems, vol. 35, pp. 36479–36494, 2022.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[7] Y. Balaji et al., "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," arXiv preprint arXiv:2211.01324, 2022.

[8] Z. Xue et al., "Raphael: Text-to-image generation via large mixture of diffusion paths," Advances in Neural Information Processing Systems, vol. 36, pp. 41693–41706, 2023.

[9] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," arXiv preprint arXiv:2208.01618, 2022.

[10] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 1931–1941.

[11] 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 22500–22510.

[12] Y. Zhang et al., "Prospect: Prompt spectrum for attribute-aware personalization of diffusion models," ACM Transactions on Graphics (TOG), vol. 42, no. 6, pp. 1–14, 2023.

[13] H. Zhang et al., "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5907–5915.

[14] T. Xu et al., "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1316–1324.

[15] 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.

[16] 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15943–15953.

[17] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," arXiv preprint arXiv:2208.01626, 2022.

[18] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," arXiv preprint arXiv:2210.11427, 2022.

[19] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 18392–18402.

[20] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2live: Text-driven layered image and video editing," in European conference on computer vision, 2022, pp. 707–723.

[21] C. Luo, "Understanding diffusion models: A unified perspective," arXiv preprint arXiv:2208.11970, 2022.

[22] X. Wang, S. Fu, Q. Huang, W. He, and H. Jiang, "Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance," arXiv preprint arXiv:2406.07209, 2024.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[24] A. Radford et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning, 2021, pp. 8748–8763.

[25] A. Van Den Oord, O. Vinyals, and others, "Neural discrete representation learning," Advances in neural information processing systems, vol. 30, 2017.

[26] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[27] D. Podell et al., "Sdxl: Improving latent diffusion models for high-resolution image synthesis," arXiv preprint arXiv:2307.01952, 2023.

[28] D. Li, J. Li, and S. Hoi, "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," Advances in Neural Information Processing Systems, vol. 36, pp. 30146–30166, 2023.

[29] J. Yu et al., "Scaling autoregressive models for content-rich text-to-image generation," arXiv preprint arXiv:2206.10789, vol. 2, no. 3, p. 5, 2022.

[30] OpenAI. 2022. ChatGPT. https://chat.openai.com/. Accessed: 2023-10-15.