ORACLE: Leveraging Mutual Information for Consistent Character Generation with LoRAs in Diffusion Models

Kiymet Akdemir, Pinar Yanardag
Department of Computer Science, Virginia Tech
Blacksburg, Virginia, USA
{kiymet, pinary}@vt.edu

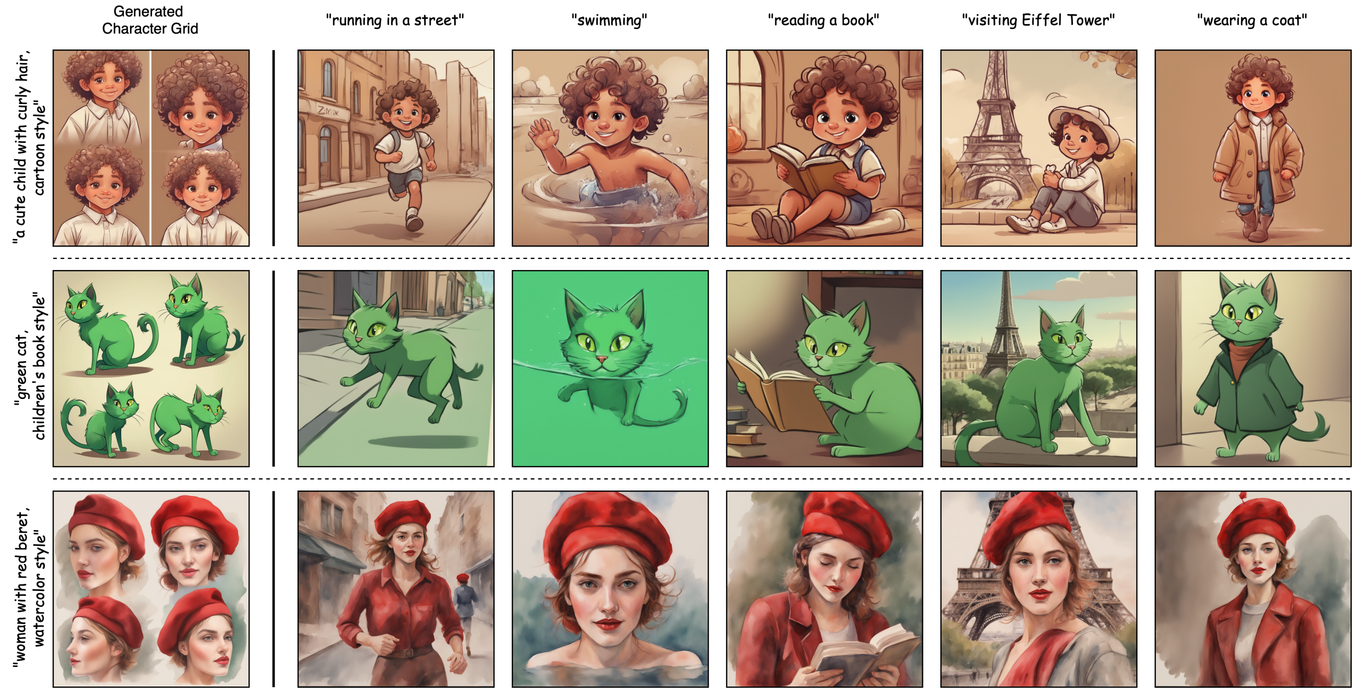
Generated Character Grid
"running in a street"
"swimming"
"reading a book"
"visiting Eiffel Tower"
"wearing a coat"
"a cute child with curly hair, cartoon style"
"green cat, children's book style"
"woman with red beret, watercolor style"

Figure 1: Given a text prompt such as 'a cute child with curly chair, cartoon style' (refer to the top row), our approach seamlessly produces consistent characters in a zero-shot manner by leveraging a pre-trained Stable Diffusion model. It ensures character consistency across a wide array of settings and backgrounds, demonstrating the versatility and practicality of our method. Our method has the potential to enhance creative process in art and design, enabling more detailed storytelling and consistent character portrayal in animations, video games, and interactive media.

Abstract

Text-to-image diffusion models have recently taken center stage as pivotal tools in promoting visual creativity across an array of domains such as comic book artistry, children's literature, game development, and web design. These models harness the power of artificial intelligence to convert textual descriptions into vivid images, thereby enabling artists and creators to bring their imaginative concepts to life with unprecedented ease. However, one of the significant hurdles that persist is the challenge of maintaining consistency in character generation across diverse contexts. Variations in textual prompts, even if minor, can yield vastly different visual outputs, posing a considerable problem in projects that require a uniform representation of characters throughout. In this paper, we introduce a novel framework designed to produce consistent character representations from a single text prompt across diverse settings. Through both quantitative and qualitative analyses, we demonstrate that our framework outperforms existing methods in generating characters with consistent visual identities, underscoring its potential to transform creative industries. By addressing the critical challenge of character consistency, we not only enhance the practical utility of these models but also broaden the horizons for artistic and creative expression.

Introduction

Text-to-image diffusion models have captivated the creative world with their extraordinary capacity to turn textual descriptions into detailed, high-resolution images. These advancements have ushered in a new era of creativity, allowing for the generation of bespoke illustrations for storybooks, dynamic characters in video games, personalized content across digital platforms, and engaging visuals for educational purposes. The ability to generate images that closely align with specific text prompts has opened up endless possibilities for storytellers, educators, game developers, and digital content creators, enabling them to bring their unique visions to life with precision and flair.

However, the journey of integrating these models into creative workflows has encountered a significant challenge: maintaining visual consistency across different scenarios. When characters are depicted in various contexts or settings, slight alterations in text prompts can lead to inconsistencies in their appearance, disrupting the visual continuity that is crucial for storytelling, brand identity, and character recognition. This challenge has been a bottleneck, limiting the full exploitation of text-to-image diffusion models in projects requiring a cohesive character narrative.

In addressing the challenge of achieving consistent character visualization across various applications of text-to-image diffusion models, the field has increasingly leaned on personalization techniques such as Dreambooth (Ruiz et al. 2022), Textual Inversion (Gal et al. 2022) or LoRAs (Hu et al. 2021). Historically, these approaches have relied extensively on reference images for character creation, a dependency that constrains their applicability across a wider range of uses. Efforts to bypass these limitations have included strategies such as manual filtering and clustering (Avrahami et al. 2023), or even the incorporation of celebrity names into prompts to guide the image synthesis process. However, such methods are typically either labor-intensive, time-consuming, or significantly restrict the diversity of characters that can be effectively rendered.

Our paper addresses this pivotal issue by presenting a novel approach that ensures the consistent generation of characters across diverse settings with a single text prompt. Based on a text prompt, for example, *"a cute child with curly hair, cartoon style"* (refer to Figure 1), our method produces a set of initial character images in a zero-shot fashion through a pre-trained text-to-image diffusion model like Stable Diffusion (Rombach et al. 2022). This candidate set undergoes refinement through a mutual information-based filtering process which then serves as the foundation for training a personalization model, such as LoRA (Hu et al. 2021). Following this process, it becomes possible to create characters that maintain consistency in a variety of settings, including diverse environments, backgrounds, and actions.

By enhancing the ability of diffusion models to maintain visual continuity, our methodology not only solves a technical problem but also profoundly impacts the creative process across multiple domains. For comic book artists and children's book authors, this breakthrough means characters can now retain their identity across panels or pages without the exhaustive effort of manual adjustments or the need for numerous reference images. This consistency is vital for narrative coherence and character development, allowing creators to focus on storytelling rather than technical limitations. In the realm of game development, our approach enables designers to create more immersive worlds, with characters that remain true to their original design throughout various game environments and scenarios. This consistency enhances the player's connection to the character and the overall gaming experience, allowing for a deeper engagement with the story. For educators and creators of educational content, this technology offers the potential to produce a wide range of consistent visual materials that can support learning objectives. Characters that recur in various educational scenarios can become memorable figures for students, aiding in engagement and the retention of information.

Our contributions are as follows:

- We propose an effective framework for producing characters that remain visually consistent across various scenarios. Our method operates in a zero-shot manner, generating unique characters that match the provided text prompts. It also eliminates discrepancies among image components using mutual information, ensuring cohesive visual representations by refining the initial set of generated images.
- We provide comprehensive qualitative and quantitative comparisons with existing methods, along with insights from a user study, highlighting the effectiveness and improvements our method provides over traditional techniques.
- The versatility and applicability of our method are highlighted through demonstrations of its use in various creative and practical contexts. We showcase how our approach enables the generation of characters that are not only consistent in appearance but also adaptable to different environments, backgrounds, and narratives, thereby broadening the potential for innovative applications in storytelling, gaming, education, and beyond.
- Additionally, we illustrate how our approach can be utilized to design compelling storylines with a story example, and transform our characters into 3D objects for gaming purposes. This demonstrates our method's ability to not only create visually consistent characters but also to support the broader creative processes involved in narrative development and interactive game design.

These contributions collectively enhance the toolbox available to creators and developers, offering new pathways to leverage text-to-image diffusion models for creating coherent and engaging visual narratives.

Related Work

Text-to-image Generation

The advancement of large-scale text-to-image diffusion models (Rombach et al. 2022; Nichol et al. 2021) has enabled the widespread use of image generation models (Chen et al. 2020; Kim, Son, and Kim 2021; Saharia et al. 2022), largely due to the plentiful availability of image-text pair datasets and their simpler training process when compared to Generative Adversarial Networks (GANs) (Goodfellow et al. 2014). While these models excel in producing a wide array of realistic images, they struggle with generating consistent images; minor changes in the prompt can lead to substantial variations in the outputs. This limitation largely restricts their applicability for illustration purposes, where consistency is key.

Text-to-image Personalization

Current text-to-image models struggle to depict specific entities across varied contexts. However, advancements in personalization techniques now allow for the creation of images depicting a given subject in various contexts utilizing a reference image set. Textual Inversion (Gal et al. 2022) aims to integrate particular instances or styles into the embedding space of existing, unmodified text-to-image models. DreamBooth (Ruiz et al. 2022), on the other hand, suggests fine-tuning the entire model to associate specific instances with a unique identifier, while still maintaining the instance's general category. This approach, though, necessitates saving a separate model for each subject, leading to storage issues. LoRA (Hu et al. 2021) enables the fine-tuning of a limited number of parameters, significantly simplifying the storage of numerous personalized models. One other technique used for personalization purposes, the IP Adapter (Ye et al. 2023), employs an image encoder to integrate image features into the diffusion model. Nevertheless, it struggles accurately adhering to text prompts across varying contexts. These techniques are limited by the need for a reference image set provided by the user, which constrains the diversity and creativity of the subjects illustrated.

Generating Consistent Characters

Recent attempts in story illustration face limitations, as they are trained on specialized datasets (Rahman et al. 2023), depend on personalization models that require a set of reference images (Gong et al. 2023), or utilize face-swapping techniques (Jeong, Kwon, and Ye 2023), which confines their use of human subjects. Consequently, these story illustration methods often limit character representations to pre-existing entities. Alternative approaches involve either manual image filtering or incorporating celebrity names into prompts, with the former being time-consuming and the latter narrowly confining the range of potential subjects for illustrations. The Chosen One (Avrahami et al. 2023) touches on the problem of generating imaginative characters and suggests generating and clustering images based on a specific prompt, then using the most consistent image cluster to iteratively train a personalized model until it reaches convergence. Yet, it demands considerable time for image generation and multiple rounds of model training.

Background

Diffusion models

Diffusion models are a class of generative models that estimates the complex data distribution through iterative denoising process. As the source of diversity, the initial latent $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is sampled and fed to U-Net model, ϵ_θ, as an attempt to gradually denoise the latent variable x_t to obtain x_{t-1} for T timesteps where x_1 corresponds to the final image. The joint probability of latent variables $\{x_1, ..., x_T\}$ is modeled as a Markov Chain.

$$p_\theta(x_{1:T}) = p(x_T) \prod_{t=T}^{1} p_\theta(x_{t-1}|x_t) \tag{1}$$

In text-to-image generation task, diffusion models are conditioned on an external text input c where the overall aim is producing an image that is aligned with the description provided by c. To train text-to-image models, diffusion models use a simplified objective.

$$\mathbb{E}_{x,c,\epsilon,t}\left[\|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2\right], \tag{2}$$

where (x_t, c) is latent-text condition pair, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \sim \mathcal{U}([0, 1])$. For unconditional generation, c is set to null text. In inference stage, classifier free guidance is applied to noise prediction to improve the sample quality.

$$\tilde{\epsilon}_\theta(x_t, t, c) = \epsilon_\theta(x_t, t, \emptyset) + \gamma[\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \emptyset)], \tag{3}$$

where $\gamma \geq 1$ is the guidance scale and $\emptyset$ denotes the null text condition.

Mutual Information

Using mutual information in the field of computer vision, first proposed by (Veale and Hao 2007; Maes et al. 1997) and has been employed as a robust method for comparing image similarity (Klein, Staring, and Pluim 2007; Kothandaraman et al. 2023) by binning pixel values into histograms and comparing their distributions. Mutual information serves as a metric to measure the information acquired about one variable upon observing another variable. It effectively captures the dependence between two random variables, X and Y, using the following equation:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{4}$$

where

$$H(X) = -\sum_x P_X(x) \log P_X(x) = -E_{P_X} \log P_X, \tag{5}$$

$$\begin{aligned} H(X|Y) &= \sum_y P_Y(y) \left[-\sum_x P_{X|Y}(x|y) \log\left(P_{X|Y}(x|y)\right) \right] \\ &= E_{P_Y}\left[-E_{P_{X|Y}} \log P_{X|Y}\right] \end{aligned} \tag{6}$$

Entropy, a core concept in information theory, quantifies the level of unpredictability associated with a variable's outcomes. Here, $H(X)$ represents the entropy of X, signifying the inherent uncertainty or randomness of X, while $H(X|Y)$ denotes the conditional entropy of X given Y, which measures the remaining uncertainty in X once Y is known. The conditional entropy, $H(X|Y)$, specifically quantifies the extent to which the uncertainty of X is reduced by knowing the outcome of Y. The mutual information formula, therefore, quantifies the reduction in uncertainty about one variable given knowledge of the other.

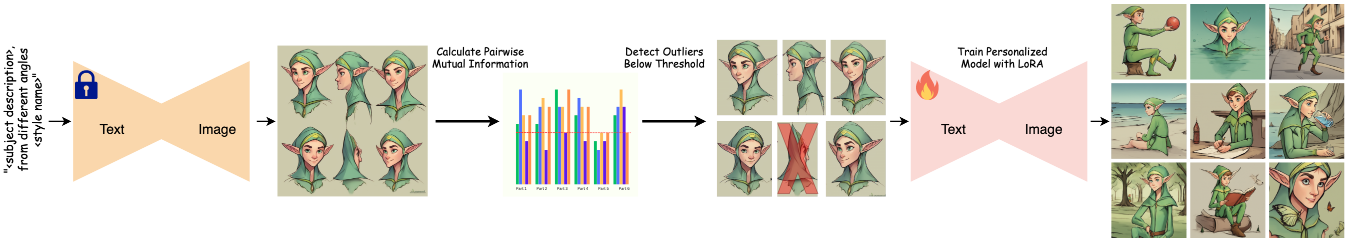
"<subject description>, from different angles <style name>"
Text
Image
Calculate Pairwise Mutual Information
Part 1
Part 2
Part 3
Part 4
Part 5
Part 6
Detect Outliers Below Threshold
Train Personalized Model with LoRA
Text
Image

Figure 2: **An overview of ORACLE.** Our method operates through three phases: 1) It begins with the generation of a grid based on structured prompts that include character description, style, and a grid generator prompt, like *"from different angles"*. 2) Subsequently, it calculates the average pairwise mutual information to identify potential outliers. 3) Once outliers are filtered out, a personalized model is trained using the refined grid segments.

Method

Our approach is structured around three key phases. The initial part involves generating a grid of images that align with the provided prompt, using the text-to-image diffusion model. The second stage focuses on identifying and removing any outliers that do not match the consistency of the rest of the images. Lastly, we tailor a model using personalization techniques specifically designed to generate images across various contexts while maintaining the details in the set of refined consistent images. An overview of our framework is given in Figure 2.

Candidate Set Generation

Given a text prompt, the initial step in our method involves generating a single grid of candidate images that correspond to the described character. In contrast to traditional approaches that depend on costly techniques such as clustering a large number of images (Avrahami et al. 2023), or the need for manually curated datasets to train a personalized model (Ruiz et al. 2022; Gal et al. 2022; Hu et al. 2021), our technique employs the "grid trick" to generate an initial set of candidates. This strategy, also referred to as a *character sheet*, has become popular within the Stable Diffusion art community, especially among enthusiasts and professionals for tasks like avatar creation and image stylization[1]. The trick involves leveraging a pretrained text-to-image model with specific directions, such as *"<character description> from multiple angles, <style description>"* or *"<character description> from different perspectives, <style description>"*. While a template grid combined with ControlNet can be used to automatically crop image parts, we found that using a template compromises from the quality and creativity of the generated characters. Therefore, we prefer to manually crop the image parts, typically involving only 4-6 sections in a character grid. Previous research such as (Kara et al. 2023) have applied this technique for video editing and highlighted its ability to generate multiple images with a consistent style is due to the diffusion model treating the grid as a singular, composite image.

Candidate Set Refinement

While the initial batch of images accurately reflects the text prompt and achieves a level of consistency, it also displays noticeable inconsistencies, such as imprecise details or significant variations (illustrated by the leftmost set of images in Fig. 2). Therefore, we employ a mutual information-based strategy to identify and remove elements that could disrupt the uniformity of the personalized model. We argue that traditional vector similarity metrics, such as cosine similarity, fall short of our needs because they tend to interpret different views of the same subject as distinct features. In contrast, mutual information proves to be well-suited for our objective by assessing the distribution of image features, offering a more nuanced and effective means of evaluating consistency across various representations.

Our objective is to generate a consistent set of images such that the average pairwise mutual information for each image within the set, i.e. $S_i = \frac{1}{k}\sum_{j=1}^{k} I(V_i, V_j)$ surpasses the predefined threshold. To achieve this, we have a binary function $\mathbf{C}$ that determines whether a specific image V_i qualifies to be part of the final collection or not:

$$\mathbf{C}(V_i) = \begin{cases} 1 & S_i \geq \mu - k\sigma \\ 0 & S_i < \mu - k\sigma \end{cases} \tag{7}$$

where μ represents the mean and σ represents the standard deviation of average pairwise mutual information for each part, and k is a strictness constant. By this binary function, we automatically eliminate the outlier components to reach an ideal mix, where each piece is unique but also fits well with the others. Note that as the constant k decreases, the filter becomes more strict.

Personalization of the Character

Lastly, we train a LoRA (Hu et al. 2021) model with DreamBooth (Ruiz et al. 2022) on the refined set of images in order to generate images across various contexts while maintaining the details of the character.

Experiments

We evaluate our method against various baselines through both quantitative and qualitative analysis. Following this, we

[1] How To Create Consistent Characters In Midjourney : https://shorturl.at/jwAJW

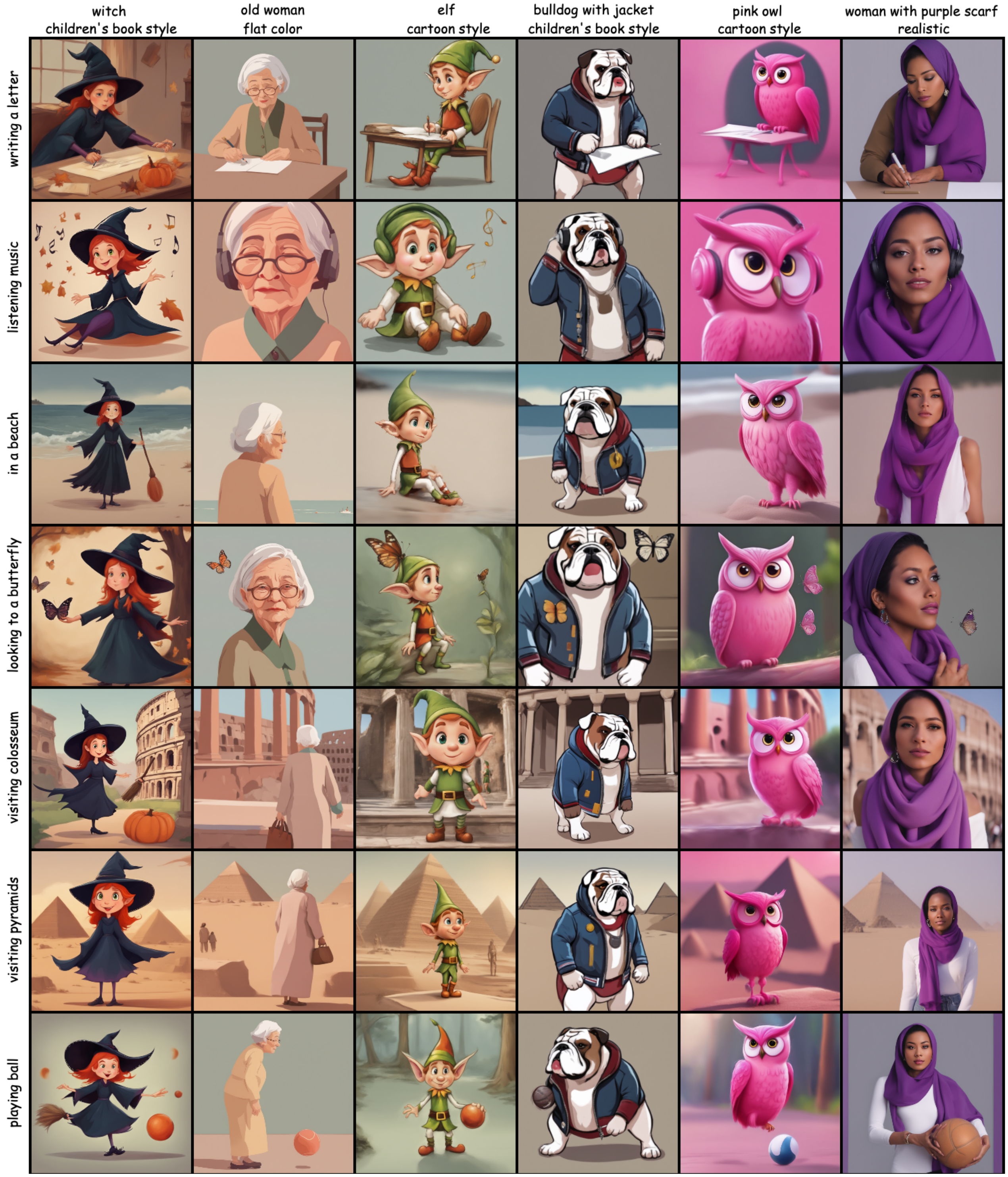

Figure 3: **Qualitative results.** Our method can produce a wide array of characters in diverse contexts and styles, from imaginative figures like *'a bulldog wearing a jacket'* and *'a pink owl'*, to photo-realistic characters such as *'a woman with a purple scarf'*.

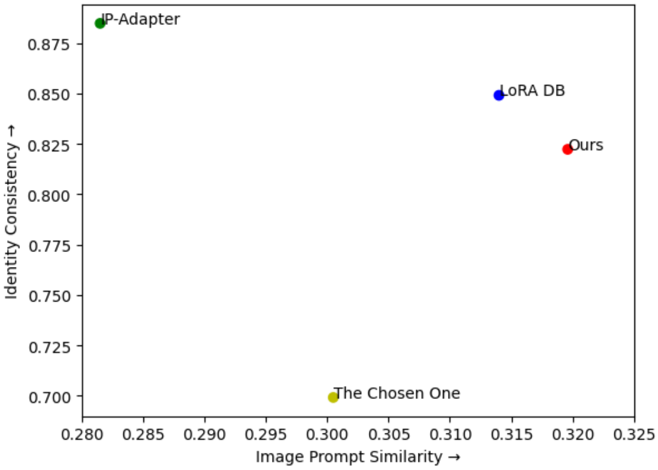

Figure 4: **Quantitative comparisons.** We use CLIP to assess the relevance of images to their prompts (image-prompt similarity) and identity consistency (image-image similarity).

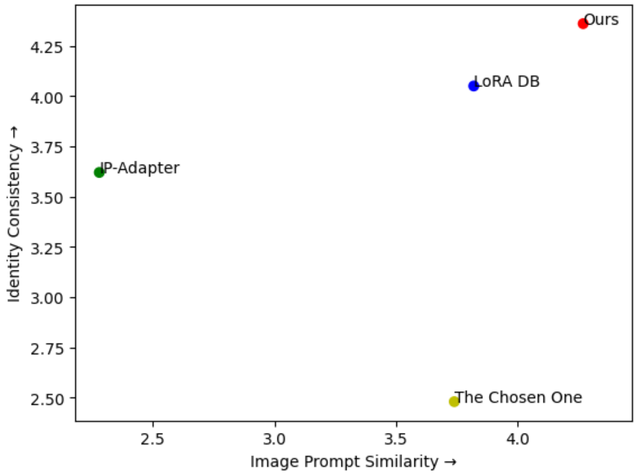

Figure 5: **User study results.** The average user rating for each baseline is given for two types of questions (identity consistency and relevance to prompt). Rating is performed on a scale from 1 to 5.

detail the findings from our user study. Lastly, we demonstrate several applications of our approach, including story illustration, object generation, and 3D reconstruction.

Baselines: We compared our model with three state-of-the-art models using SDXL (Podell et al. 2023): The Chosen One (Avrahami et al. 2023), IP-Adapter (Ye et al. 2023), and DreamBooth (Ruiz et al. 2022) combined with LoRA (Hu et al. 2021) (LoRA DB). For each method, we used the same text prompt $\mathcal{P}$ such as *'A golden dog with red hat, watercolor style'*. We use the same single image generated using prompt $\mathcal{P}$ for encoding-based models (IP-Adapter (Ye et al. 2023)), or methods that require a reference image set (LoRA DB (Ruiz et al. 2022)). In our approach, we employ a refined set of images as outlined in our methodology.

Implementation details: We used the official codebases for all competitors. All methods use a recent state-of-the-art text-to-image model, SDXL (Podell et al. 2023). We run our experiments on a single A40 Nvidia GPU. Our method requires 30 seconds to generate the candidate set, and an additional 10 seconds to calculate mutual information for refinement. In contrast, methods such as (Avrahami et al. 2023) takes 20 minutes per iteration. We used the strictness constant k in Eq. 7 for outlier elimination as 1.

Qualitative Experiments

In Figure 3, we showcase the qualitative outcomes of our methodology. Our approach effectively produces characters across diverse contexts and styles while preserving their identity. For instance, with the text prompt ‘witch, children’s book style,’ our method not only generates a distinct character but also adeptly positions it in various scenarios such as ‘on a beach,’ ‘visiting the Colosseum,’ ‘exploring the pyramids,’ along with different activities like ‘writing a letter,’ ‘listening to music,’ ‘watching a butterfly,’ or ‘playing ball.’ Furthermore, our results demonstrate the capability to adapt these scenarios for a wide array of characters, including ‘an old woman,’ ‘an elf,’ ‘a bulldog,’ ‘a pink owl,’ and even photorealistic figures like ‘a woman with a purple scarf.’

Qualitative comparison: In Figure 6, we provide a qualitative comparison between our method and other approaches. The IP-Adapter shows difficulty in adhering to prompts, such as *”old man driving”* or *”cute child with curly hair holding an umbrella”*. On the other hand, The Chosen One succeeds creating characters within specified concepts but faces challenges in generating characters that are consistent with the given prompt, as seen in the example of the *‘golden dog with red hat.”* LoRA-DB (Ruiz et al. 2022) generally succeeds in producing consistent characters, yet it fails to accurately follow the prompt, like in the case of the *‘golden dog with red hat sleeping.”* Additionally, LoRA-DB (Ruiz et al. 2022) and IP-Adapter (Ye et al. 2023) tend to keep the pose of characters unchanged across different contexts. In contrast, our method demonstrates superior ability in both following the prompt accurately and maintaining the character consistency.

Quantitative Experiments

We perform a quantitative evaluation based on two metrics: image prompt similarity and identity consistency. These metrics are widely used in studies on personalization techniques (Ruiz et al. 2022; Gal et al. 2022) and generating consistent characters (Avrahami et al. 2023; Tewel et al. 2024). We measure the normalized cosine similarity between the image and prompt text embedding using CLIP (Radford et al. 2021) in order to evaluate the image prompt similarity. Similarly, we utilized CLIP to assess identity consistency, where we calculate the average pairwise normalized cosine similarity among the images of the

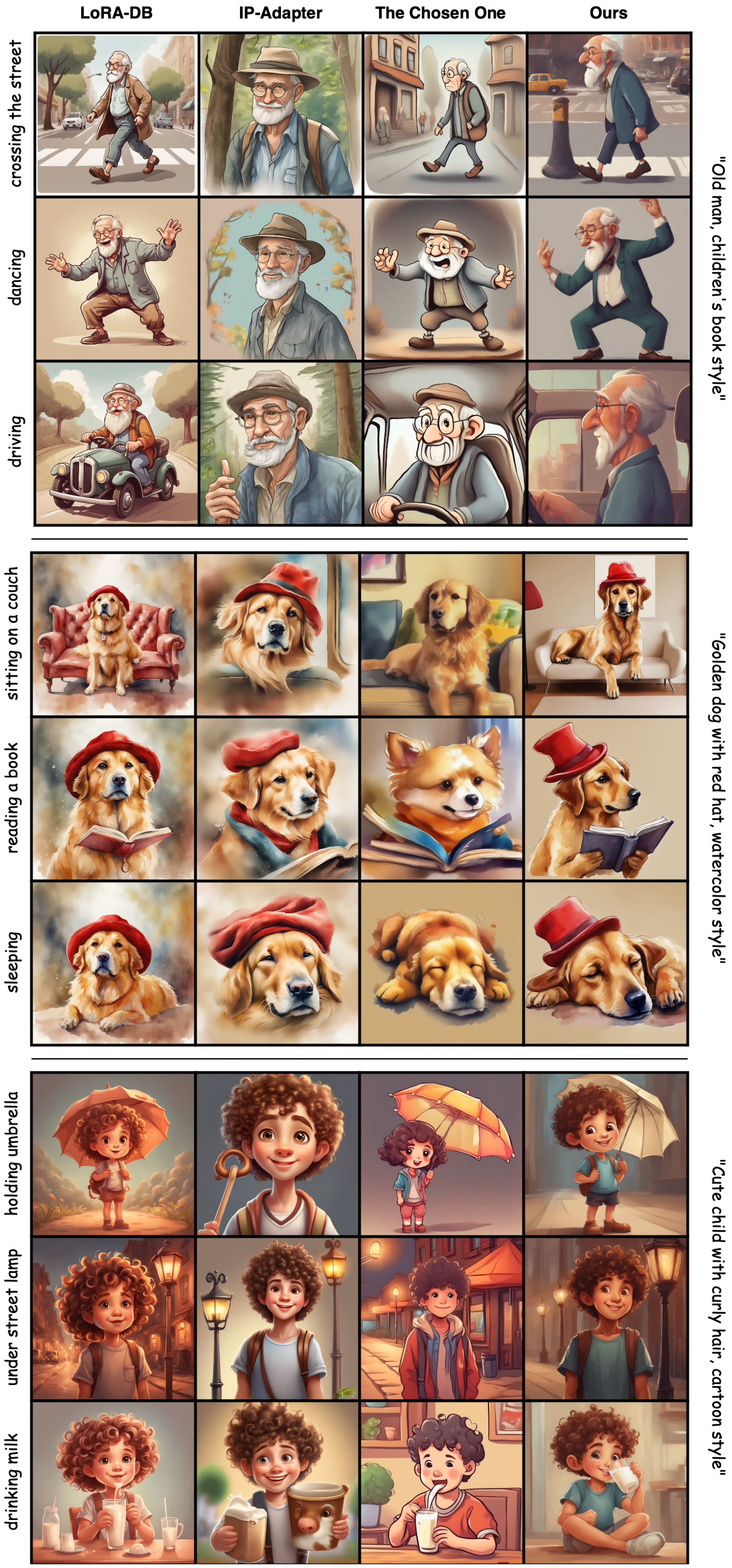

Figure 6: **Qualitative comparisons.** We compare ORACLE against various baselines, including LoRA DB, IP-Adapter, and The Chosen One. Our method surpasses these in effectively adhering to the given prompts and achieving greater consistency.

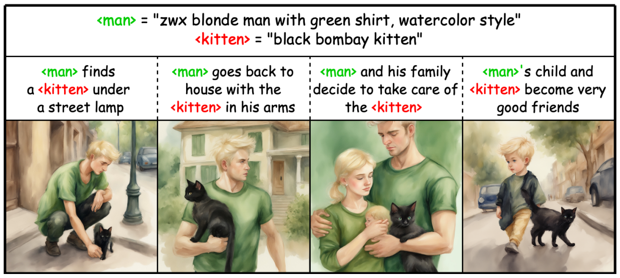
<man> = "zwx blonde man with green shirt, watercolor style"
<kitten> = "black bombay kitten"
<man> finds a <kitten> under a street lamp
<man> goes back to house with the <kitten> in his arms
<man> and his family decide to take care of the <kitten>
<man>'s child and <kitten> become very good friends

Figure 7: **Story illustration.** A demonstration of story illustration using a model trained with the specified description of the man.

same subject among different contexts. We generate 4 characters, each in 12 different contexts, using the same seeds for each method, resulting in a total of 48 images evaluated for each method.

Our quantitative findings are illustrated in Figure 4. Usually, finding a balance between preserving identity consistency and creating images that closely match the text prompt is essential. Techniques like LoRA DB (Ruiz et al. 2022) and IP-Adapter (Ye et al. 2023) perform well in preserving character identity, primarily by performing minimal changes across images, but they tend to fall short of generating images that closely follow the text prompts. On the other hand, The Chosen One (Avrahami et al. 2023) is adept at creating images that match the prompts but struggle with keeping the character's identity consistent. Our method, however, achieves an optimal balance, effectively adhering to the text prompt while ensuring the character's identity remains consistent. This quantitative analysis corroborates our qualitative observations, underscoring the effectiveness of our approach.

User Study

We carried out a user study with 54 participants via the Prolific platform. Utilizing the visuals displayed in Fig. 6, we presented a series of three images for each character, depicted in different contexts. Participants were randomly assigned a set of images and instructed to rate them on a scale from 1 to 5, evaluating both their relevance to the text prompt (image-prompt similarity) and their consistency with each other (image-image similarity). Specifically, participants were asked the following questions:

Q1: Given the text description and the images shown above, how well the images reflect the given text description? Rate from 1 (Not relevant at all) to 5 (Very Relevant)

Q2: Considering the three images presented earlier, how consistent is the character depicted across them? Rate 1 (Not consistent at all) to 5 (Very Consistent)

The average ratings for all subjects are calculated for each method and are displayed in Figure 5. Overall, the results of the user study supports the quantitative findings presented in Figure 4. Notably, our method emerged as the most preferred approach among users for both its consistency and relevance to the given text prompts.

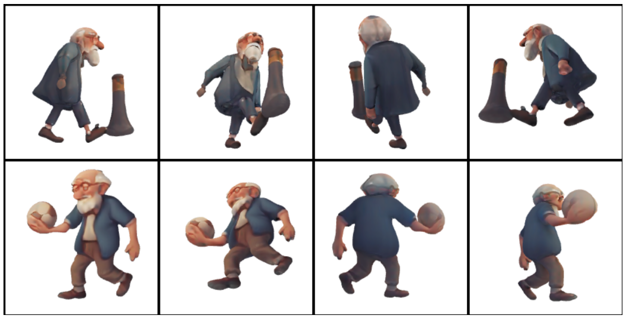

Figure 8: **An example of 3D character.**

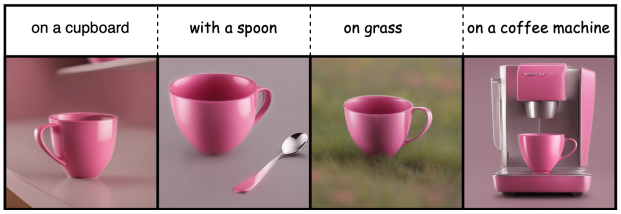

Figure 9: **An example of object generation.**

Applications

Our method have various applications, including story illustration, as demonstrated in Figure 7. Although the model is specifically trained for the man, it possesses the capability to generate images of the character's family or child, as illustrated in Figure 7. This capability broadens the scope of illustrations achievable with a single trained model. Moreover, our characters can be transformed into 3D as in Fig. 8, using off the shelf methods such as TripoSR (Tochilkin et al. 2024). Additionally, we illustrate that our method is also effective at generating consistent objects in addition to characters, as shown in Figure 9.

Limitations

Even though our method can successfully generate consistent characters, inherent limitations associated with Stable Diffusion model exists; even if the images fed into the personalized model are perfectly consistent, the model might still alter certain details, such as clothing, across different contexts—a variation that might be desirable in specific scenarios.

Conclusion

In conclusion, our work introduces a lightweight, fast, and efficient strategy for creating consistent characters through text-to-image models. Our experiments reveal that this approach successfully ensures consistency across various outputs and maintains alignment with the given text prompts, as evidenced by quantitative scores. This achievement is further validated by qualitative evaluations and a user study. Our method is opening up new avenues for utilizing text-to-image diffusion models to craft cohesive and captivating visual stories.

References

Avrahami, O.; Hertz, A.; Vinker, Y.; Arar, M.; Fruchter, S.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. The chosen one: Consistent characters in text-to-image diffusion models. *arXiv preprint arXiv:2311.10093*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

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

Gong, Y.; Pang, Y.; Cun, X.; Xia, M.; He, Y.; Chen, H.; Wang, L.; Zhang, Y.; Wang, X.; Shan, Y.; and Yang, Y. 2023. Talecrafter: Interactive story visualization with multiple characters.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models.

Jeong, H.; Kwon, G.; and Ye, J. C. 2023. Zero-shot generation of coherent storybook from plain text story using diffusion models.

Kara, O.; Kurtkaya, B.; Yesiltepe, H.; Rehg, J. M.; and Yanardag, P. 2023. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. *arXiv preprint arXiv:2312.04524*.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Klein, S.; Staring, M.; and Pluim, J. P. W. 2007. Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. *IEEE Transactions on Image Processing* 16(12):2879–2890.

Kothandaraman, D.; Zhou, T.; Lin, M.; and Manocha, D. 2023. Aerialbooth: Mutual information guidance for text controlled aerial view synthesis from a single image. *arXiv preprint arXiv:2311.15478*.

Maes, F.; Collignon, A.; Vandermeulen, D.; Marchal, G.; and Suetens, P. 1997. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging* 16(2):187–198.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision.

Rahman, T.; Lee, H.-Y.; Ren, J.; Tulyakov, S.; Mahajan, S.; and Sigal, L. 2023. Make-a-story: Visual memory conditioned consistent story generation.

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

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

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35:36479–36494.

Tewel, Y.; Kaduri, O.; Gal, R.; Kasten, Y.; Wolf, L.; Chechik, G.; and Atzmon, Y. 2024. Training-free consistent text-to-image generation.

Tochilkin, D.; Pankratz, D.; Liu, Z.; Huang, Z.; Letts, A.; Li, Y.; Liang, D.; Laforte, C.; Jampani, V.; and Cao, Y.-P. 2024. Triposr: Fast 3d object reconstruction from a single image.

Veale, T., and Hao, Y. 2007. Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07)*, 1471–1476. Vancouver, British Columbia: AAAI Press.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models.