The Chosen One: Consistent Characters in Text-to-Image Diffusion Models

Omri Avrahami*
The Hebrew University of Jerusalem
Google Research
Jerusalem, Israel
omri.avrahami@mail.huji.ac.il

Amir Hertz
Google Research
Tel Aviv, Israel
hertzamir@gmail.com

Yael Vinker*
Tel Aviv University
Google Research
Tel Aviv, Israel
yaelvi116@gmail.com

Moab Arar*
Tel Aviv University
Google Research
Tel Aviv, Israel
moab.arar@gmail.com

Shlomi Fruchter
Google Research
Tel Aviv, Israel
shlomi.fruchter@gmail.com

Ohad Fried
Reichman University
Herzliya, Israel
ofried@runi.ac.il

Daniel Cohen-Or*
Tel Aviv University
Google Research
Tel Aviv, Israel
cohenor@gmail.com

Dani Lischinski*
The Hebrew University of Jerusalem
Google Research
Jerusalem, Israel
danix@mail.huji.ac.il

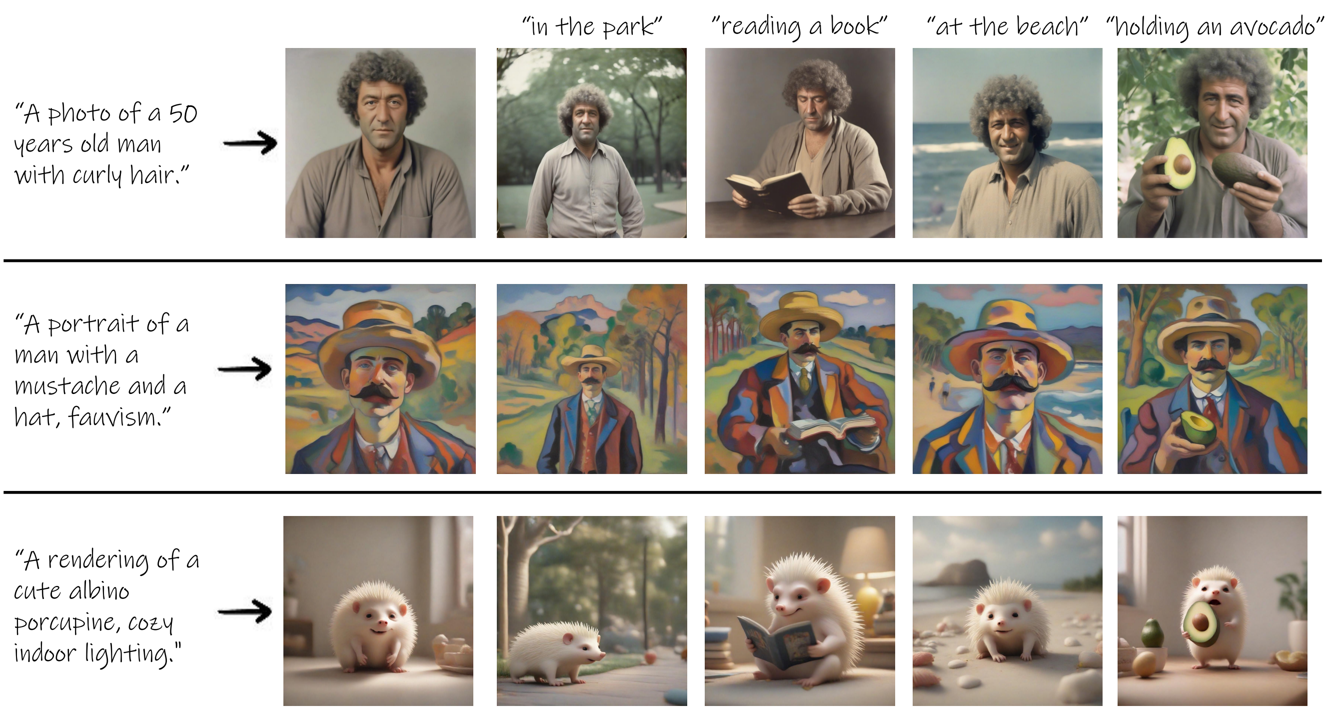


Figure 1: The Chosen One: Given a text prompt describing a character, our method distills a representation that enables consistent depiction of *the same character* in novel contexts.

*Performed this work while working at Google

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0525-0/24/07.
https://doi.org/10.1145/3641519.3657430

ABSTRACT

Recent advances in text-to-image generation models have unlocked vast potential for visual creativity. However, the users that use these models struggle with the generation of consistent characters, a crucial aspect for numerous real-world applications such as story visualization, game development, asset design, advertising, and more. Current methods typically rely on multiple pre-existing images of the target character or involve labor-intensive manual processes. In this work, we propose a fully automated solution for consistent character generation, with the sole input being a text prompt. We introduce an iterative procedure that, at each stage, identifies a coherent set of images sharing a similar identity and extracts a more consistent identity from this set. Our quantitative analysis demonstrates that our method strikes a better balance between prompt alignment and identity consistency compared to the baseline methods, and these findings are reinforced by a user study. To conclude, we showcase several practical applications of our approach.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Computer graphics**.

KEYWORDS

Consistent characters generation

ACM Reference Format:
Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2024. The Chosen One: Consistent Characters in Text-to-Image Diffusion Models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3641519.3657430

1 INTRODUCTION

The ability to maintain consistency of generated visual content across various contexts, as shown in Figure 1, plays a central role in numerous creative endeavors. These include illustrating a book, crafting a brand, creating comics, developing presentations, designing webpages, and more. Such consistency serves as the foundation for establishing brand identity, facilitating storytelling, enhancing communication, and nurturing emotional engagement.

Despite the increasingly impressive abilities of text-to-image generative models, the users that use these models struggle with such consistent generation, a shortcoming that we aim to rectify in this work. Specifically, we introduce the task of *consistent character generation*, where given an input text prompt describing a character, we derive a representation that enables generating consistent depictions of the same character in novel contexts. Although we refer to characters throughout this paper, our work is in fact applicable to visual subjects in general.

Project page is available at: https://omriavrahami.com/the-chosen-one/

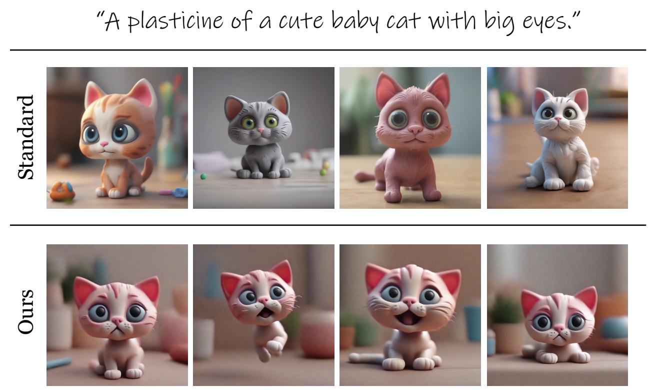


Figure 2: Identity consistency. Given the prompt *"a Plasticine of a cute baby cat with big eyes"*, a standard text-to-image diffusion model produces different cats (all corresponding to the input text), whereas our method produces the *same cat.*

Consider, for example, an illustrator working on a Plasticine cat character. As demonstrated in Figure 2, providing a state-of-the-art text-to-image model with a prompt describing the character, results in a variety of outcomes, which may lack consistency (top row). In contrast, in this work we show how to distill a consistent representation of the cat (2nd row), which can then be used to depict the *same character* in a multitude of different contexts.

The widespread popularity of text-to-image generative models [Podell et al. 2023; Ramesh et al. 2022; Rombach et al. 2021; Saharia et al. 2022], combined with the need for consistent character generation, has already spawned a variety of ad hoc solutions. These include, for example, using celebrity names in prompts [stassius 2023] for creating consistent humans, or using image variations [Ramesh et al. 2022] and filtering them manually by similarity [JoshGreat 2023]. In contrast to these ad hoc, manually intensive solutions, we propose a fully-automatic principled approach to consistent character generation.

The academic works most closely related to our setting are ones dealing with personalization [Gal et al. 2022; Ruiz et al. 2023] and story generation [Gong et al. 2023; Jeong et al. 2023; Rahman et al. 2022]. Some of these methods derive a representation for a given character from *several* user-provided images [Gal et al. 2022; Gong et al. 2023; Ruiz et al. 2023]. Others cannot generalize to novel characters that are not in the training data [Rahman et al. 2022], or rely on textual inversion of an existing depiction of a human face [Jeong et al. 2023].

In this work, we argue that in many applications the goal is to generate *some* consistent character, rather than visually matching a specific appearance. Thus, we address a new setting, where we aim to automatically distill a consistent representation of a character that is only required to comply with a single natural language description. Our method does not require *any* images of the target character as input; thus, it enables creating a *novel* consistent character that does not necessarily resemble any existing visual depiction.

Our fully-automated solution to the task of consistent character generation is based on the assumption that a sufficiently large set of generated images, for a certain prompt, will contain groups of images with shared characteristics. Given such a cluster, one can extract a representation that captures the "common ground" among its images. Repeating the process with this representation, we can increase the consistency among the generated images, while still remaining faithful to the original input prompt.

We start by generating a gallery of images based on the provided text prompt, and embed them in a Euclidean space using a pre-trained feature extractor. Next, we cluster these embeddings, and choose the most *cohesive* cluster to serve as the input for a personalization method that attempts to extract a consistent identity. We then use the resulting model to generate the next gallery of images, which should exhibit more consistency, while still depicting the input prompt. This process is repeated iteratively until convergence.

We evaluate our method quantitatively and qualitatively against several baselines, as well as conducting a user study. Finally, we present several applications of our method.

In summary, our contributions are: (1) we formalize the task of consistent character generation, (2) propose a novel solution to this task, and (3) we evaluate our method quantitatively and qualitatively, in addition to a user study, to demonstrate its effectiveness.

2 RELATED WORK

Text-to-image generation. Text conditioned image generative models (T2I) [Ramesh et al. 2022; Rombach et al. 2021; Yu et al. 2022] show unprecedented capabilities of generating high quality images from mere natural language text descriptions. They are quickly becoming a fundamental tool for any creative vision task. In particular, text-to-image diffusion models [Balaji et al. 2022; Ho et al. 2020; Nichol et al. 2021; Sohl-Dickstein et al. 2015; Song et al. 2020; Song and Ermon 2019] are employed for guided image synthesis [Avrahami et al. 2023c; Chefer et al. 2023; Couairon et al. 2023; Ge et al. 2023; Hertz et al. 2022; Mou et al. 2023; Voynov et al. 2022; Zhang et al. 2023a] and image editing tasks [Avrahami et al. 2023b, 2022; Bar-Tal et al. 2022; Cao et al. 2023; Hertz et al. 2023; Kawar et al. 2023; Meng et al. 2021; Mokady et al. 2023; Patashnik et al. 2023; Sheynin et al. 2022; Tumanyan et al. 2023]. Using image editing methods, one can edit an image of a given character, and change its pose, etc., however, these methods cannot ensure consistency of the character in novel contexts, as our problem dictates.

In addition, diffusion models were used in other tasks [Po et al. 2023; Zhang et al. 2023b], such as: video editing [Geyer et al. 2023; Liu et al. 2023a,b; Molad et al. 2023; Qi et al. 2023; Yang et al. 2023], 3D synthesis [Fridman et al. 2023; Höllein et al. 2023; Metzer et al. 2023; Poole et al. 2022], editing [Benaim et al. 2022; Gordon et al. 2023; Sella et al. 2023; Zhuang et al. 2023] and texturing [Richardson et al. 2023b], typography generation [Iluz et al. 2023], motion generation [Raab et al. 2023; Tevet et al. 2022], and solving inverse problems [Horwitz and Hoshen 2022].

Text-to-image personalization. Text-conditioned models cannot generate an image of a specific object or character. To overcome this limitation, a line of works utilizes *several* images of the same instance to encapsulate new priors in the generative model. Existing solutions range from optimization of text tokens [Gal et al. 2022; Vinker et al. 2023; Voynov et al. 2023] to fine-tuning the parameters of the entire model [Avrahami et al. 2023a; Ruiz et al. 2023], where in the middle, recent works suggest fine-tuning a small subset of parameters [Alaluf et al. 2023; Chen et al. 2023b; Han et al. 2023; Hu et al. 2021; Kumari et al. 2023; Ryu 2022; Tewel et al. 2023]. Models trained in this manner can generate consistent images of the same subject. However, they typically require a *collection* of images depicting the subject, which naturally narrows their ability to generate any imaginary character. Moreover, when training on a single input image [Avrahami et al. 2023a], these methods tend to overfit and produce similar images with minimal diversity during inference.

Unlike previous works, our method does not require an input image; instead, it can generate consistent and diverse images of the same character based only on a text description. Additional works are aimed to bypass the personalization training by introducing a dedicated personalization encoder [Arar et al. 2023; Chen et al. 2023a; Gal et al. 2023; Jia et al. 2023; Li et al. 2023; Shi et al. 2023; Valevski et al. 2023; Wei et al. 2023; Ye et al. 2023]. Given an image and a prompt, these works can produce images with a character similar to the input. However, as shown in Section 4.1, they lack consistency when generating multiple images from the same input. Concurrently, ConceptLab [Richardson et al. 2023a] is able to generate new members of a broad *category* (e.g., a new pet); in contrast, we seek a consistent *instance* of a character described by the input text prompt. Another line of works, focuses on learning styles [Ahn et al. 2023; Sohn et al. 2023] from a reference image. On the other hand, our work focuses on generating novel consistent characters rather than styles.

Story visualization. Consistent character generation is well studied in the field of story visualization. Early GAN works [Li et al. 2019; Szűcs and Al-Shouha 2022] employ a story discriminator for the image-text alignment. Recent works, such as StoryDALL-E [Maharana et al. 2022] and Make-A-Story [Rahman et al. 2022] utilize pre-trained T2I models for the image generation, while an adapter model is trained to embed story captions and previous images into the T2I model. However, those methods cannot generalize to novel characters, as they are trained over specific datasets. More closely related, Jeong et al. [Jeong et al. 2023] generate consistent storybooks by combining textual inversion with a face-swapping mechanism; therefore, their work relies on images of existing human-like characters. TaleCrafter [Gong et al. 2023] presents a comprehensive pipeline for storybook visualization. However, their consistent character module is based on an existing personalization method that requires fine-tuning on *several* images of the same character.

Manual methods. Other attempts for achieving consistent character generation using a generative model rely on ad hoc and manually-intensive tricks such as using text tokens of a celebrity, or a combination of celebrities [stassius 2023] in order to create a consistent human; however, the generated characters resemble the original celebrities, and this approach does not generalize to other character types (e.g., animals). Users have also proposed to ensure consistency by manually crafting very long and elaborate text prompts [JoshGreat 2023], or by using image variations [Ramesh et al. 2022] and filtering them manually by similarity [JoshGreat 2023]. Other users suggested generating a full design sheet of a

ALGORITHM 1: Consistent Character Generation

Input: Text-to-image diffusion model M, parameterized by $\Theta = (\theta, \tau)$, where θ are the LoRA weights and τ is a set of custom text embeddings, target prompt p, feature extractor F.
Hyper-parameters: number of generated images per step N, minimum cluster size $d_{min\text{-}c}$, target cluster size $d_{size\text{-}c}$, convergence criterion d_{conv}, maximum number of iterations d_{iter}
Output: a consistent representation $\Theta(p)$

repeat
$S = \bigcup_N F(M_\Theta(p))$
$C = \text{K-MEANS++}(S, k = \lfloor N/d_{size\text{-}c} \rfloor)$
$C = \{c \in C | d_{min\text{-}c} < |c|\}$ {filter small clusters}
$c_{cohesive} = \underset{c \in C}{\text{argmin}} \frac{1}{|c|} \sum_{e \in c} \|e - c_{cen}\|^2$
$\Theta = \underset{(\theta,\tau)}{\text{argmin}} \mathcal{L}_{rec}$ over $c_{cohesive}$
until $d_{conv} \geq \frac{1}{|S|^2} \sum_{s_1, s_2 \in S} \|s_1 - s_2\|^2$
return Θ

character, then manually filter the best results and use them for further generation [Foundations 2023]. All these methods are manual, labor-intensive, and ad hoc for specific domains (e.g., humans). In contrast, our method is fully automated and domain-agnostic.

3 METHOD

As stated earlier, our goal in this work is to enable generation of consistent images of a character (or another kind of visual subject) based on a textual description. We achieve this by iteratively customizing a pre-trained text-to-image model, using sets of images generated by the model itself as training data. Intuitively, we refine the representation of the target character by repeatedly funneling the model's output into a consistent identity. Once the process has converged, the resulting model can be used to generate consistent images of the target character in novel contexts. In this section, we describe our method in detail.

Formally, we are given a text-to-image model M_Θ, parameterized by Θ, and a text prompt p that describes a target character. The parameters Θ consist of a set of model weights θ and a set of custom text embeddings τ. We seek a representation $\Theta(p)$, s.t., the parameterized model $M_{\Theta(p)}$ is able to generate consistent images of the character described by p in novel contexts.

Our approach, described in Algorithm 1 and depicted in Figure 3, is based on the premise that a sufficiently large set of images generated by M for the same text prompt, but with different seeds, will reflect the non-uniform density of the manifold of generated images. Specifically, we expect to find some groups of images with shared characteristics. The "common ground" among the images in one of these groups can be used to refine the representation $\Theta(p)$ so as to better capture and fit the target. We therefore propose to iteratively cluster the generated images, and use the most cohesive cluster to refine $\Theta(p)$. This process is repeated, with the refined representation $\Theta(p)$, until convergence. Below, we describe the clustering and the representation refinement components of our method in detail.

3.1 Identity Clustering

We start each iteration by using M_Θ, parameterized with the current representation Θ, to generate a collection of N images, each corresponding to a different random seed. Each image is embedded in a high-dimensional semantic embedding space, using a feature extractor F, to form a set of embeddings $S = \bigcup_N F(M_\Theta(p))$. In our experiments, we use DINOv2 [Oquab et al. 2023] as the feature extractor F.

Next, we use the K-MEANS++ [Arthur and Vassilvitskii 2007] algorithm to cluster the embeddings of the generated images according to cosine similarity in the embedding space. We filter the resulting collection of clusters C by removing all clusters whose size is below a pre-defined threshold $d_{min\text{-}c}$, as it was shown [Avrahami et al. 2023a] that personalization algorithms are prone to overfitting on small datasets. Among the remaining clusters, we choose the most *cohesive* one to serve as the input for the identity extraction stage (see Figure 4). We define the cohesion of a cluster c as the average distance between the members of c and its centroid c_{cen}:

$$cohesion(c) = \frac{1}{|c|} \sum_{e \in c} \|e - c_{cen}\|^2. \quad (1)$$

In Figure 4 we show a visualization of the DINOv2 embedding space, where the high-dimensional embeddings S are projected into 2D using t-SNE [Hinton and Roweis 2002] and colored according to their K-MEANS++ [Arthur and Vassilvitskii 2007] clusters. Some of the embeddings are clustered together more tightly than others, and the black cluster is chosen as the most cohesive one.

3.2 Identity Extraction

Depending on the diversity of the image set generated in the current iteration, the most cohesive cluster $c_{cohesive}$ may still exhibit an inconsistent identity, as can be seen in Figure 3. The representation Θ is therefore not yet ready for consistent generation, and we further refine it by training on the images in $c_{cohesive}$ to extract a more consistent identity. This refinement is performed using text-to-image personalization methods [Gal et al. 2022; Ruiz et al. 2023], which aim to extract a character from a given set of several images that already depict a *consistent identity*. While we apply them to a set of images which are not completely consistent, the fact that these images are chosen based on their semantic similarity to each other, enables these methods to nevertheless distill a common identity from them. This way, our method can overcome the inconsistencies that may emerge due to the feature extractor F or the clustering algorithm.

We base our solution on a pre-trained Stable Diffusion XL (SDXL) [Podell et al. 2023] model, which utilizes two text encoders: CLIP [Radford et al. 2021] and OpenCLIP [Ilharco et al. 2021]. We perform textual inversion [Gal et al. 2022] to add a new pair of textual tokens τ, one for each of the two text encoders. However, we found that this parameter space is not expressive enough, as demonstrated in Section 4.3, hence we also update the model weights θ via a low-rank adaptation (LoRA) [Hu et al. 2021; Ryu 2022] of the self- and cross-attention layers of the model.

We use the standard denoising loss:

$$\mathcal{L}_{rec} = \mathbb{E}_{x \sim c_{cohesive}, z \sim E(x), \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_{\Theta(p)}(z_t, t)\|_2^2 \right], \quad (2)$$

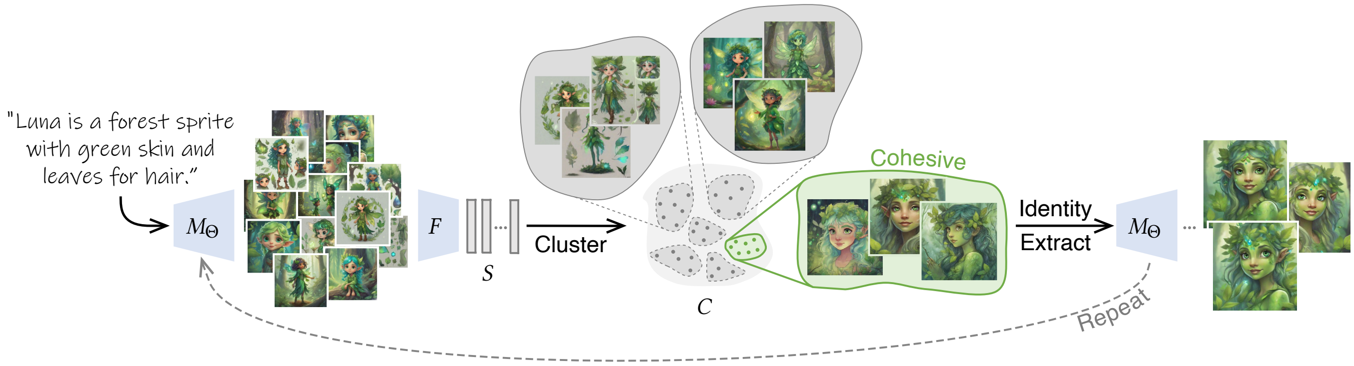


Figure 3: Method overview. Given an input text prompt, we start by generating numerous images using the text-to-image model M_Θ, which are embedded into a semantic feature space using the feature extractor F. Next, these embeddings are clustered and the most cohesive group is chosen, since it contains images with shared characteristics. The "common ground" among the images in this set is used to refine the representation Θ to better capture and fit the target. These steps are iterated until convergence to a consistent identity.

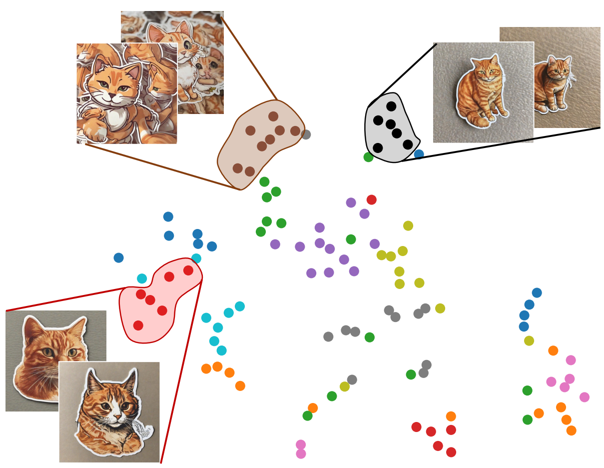


Figure 4: Embedding visualization. Given generated images for the text prompt *"a sticker of a ginger cat"*, we project the set S of their high-dimensional embeddings into 2D using t-SNE [Hinton and Roweis 2002] and indicate different K-MEANS++ [Arthur and Vassilvitskii 2007] clusters using different colors. Representative images are shown for three of the clusters. It may be seen that images in each cluster share the same characteristics: black cluster — full body cats, red cluster — cat heads, brown cluster — images with multiple cats. According to our cohesion measure (1), the black cluster is the most cohesive, and therefore, chosen for identity extraction (or refinement).

where $c_{cohesive}$ is the chosen cluster, $E(x)$ is the VAE encoder of SDXL, ϵ is the sample's noise and t is the time step, z_t is the latent z noised to time step t. We optimize $\mathcal{L}_{rec}$ over $\Theta = (\theta, \tau)$, the union of the LoRA weights and the newly-added textual tokens.

3.3 Convergence

As explained earlier (Algorithm 1 and Figure 3), the above process is performed iteratively. Note that the representation Θ extracted in each iteration is the one used to generate the set of N images for the next iteration. The generated images are thus funneled into a consistent identity.

Rather than using a fixed number of iterations, we apply a convergence criterion that enables early stopping. After each iteration, we calculate the average pairwise Euclidean distance between all N embeddings of the newly-generated images, and stop when this distance is smaller than a pre-defined threshold d_{conv}.

Finally, it should be noticed that our method is non-deterministic, i.e., when running our method multiple times, on the same input prompt p, different consistent characters will be generated. This is aligned with the one-to-many nature of our task. For more details and examples, please refer to the supplementary material.

4 EXPERIMENTS

In Section 4.1 we compare our method against several baselines, both qualitatively and quantitatively. Next, in Section 4.2 we describe the user study we conducted and present its results. The results of an ablation study are reported in Section 4.3. Finally, in Section 4.4 we demonstrate several applications of our method.

4.1 Qualitative and Quantitative Comparison

We compared our method against the most related personalization techniques [Gal et al. 2022; Li et al. 2023; Ryu 2022; Wei 2023; Ye et al. 2023]. In each experiment, each of these techniques is used to extract a character from a single image, generated by SDXL [Podell et al. 2023] from an input prompt p. The same prompt p is also provided as input to our method. Textual Inversion (TI) [Gal et al. 2022] optimizes a textual token using several images of the same concept, and we converted it to support SDXL by learning *two* text tokens, one for each of its text encoders, as we did in our method. In addition, we used LoRA DreamBooth [Ryu 2022] (LoRA DB), which we found less prone to overfitting than standard

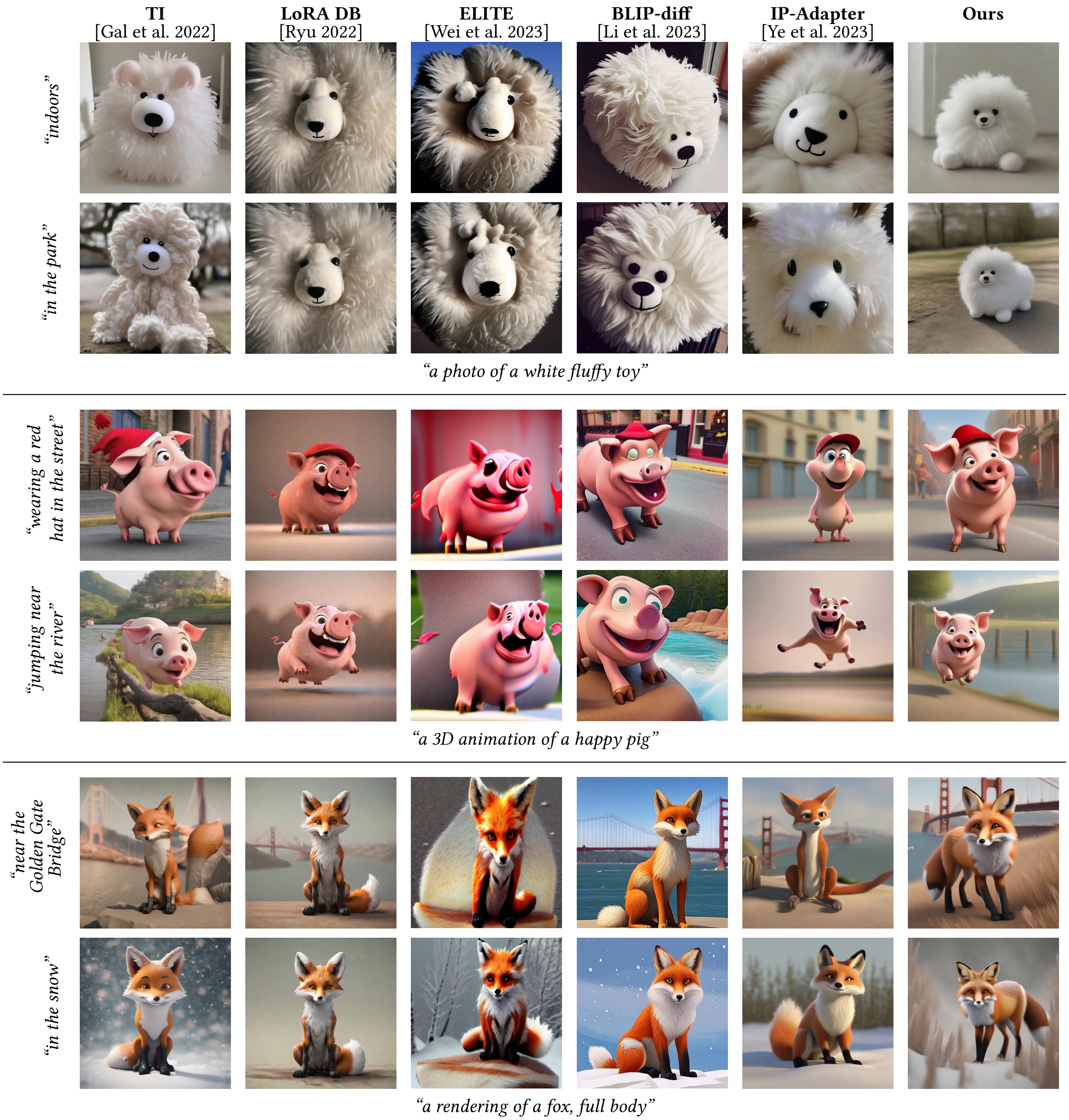


Figure 5: Qualitative comparison. We compare our method against several baselines: TI [Gal et al. 2022], BLIP-diffusion [Li et al. 2023] and IP-adapter [Ye et al. 2023] are able to follow the target prompts, but do not preserve a consistent identity. LoRA DB [Ryu 2022] is able to maintain consistency, but it does not always follow the prompt. Furthermore, the character is generated in the same fixed pose. ELITE [Wei et al. 2023] 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.

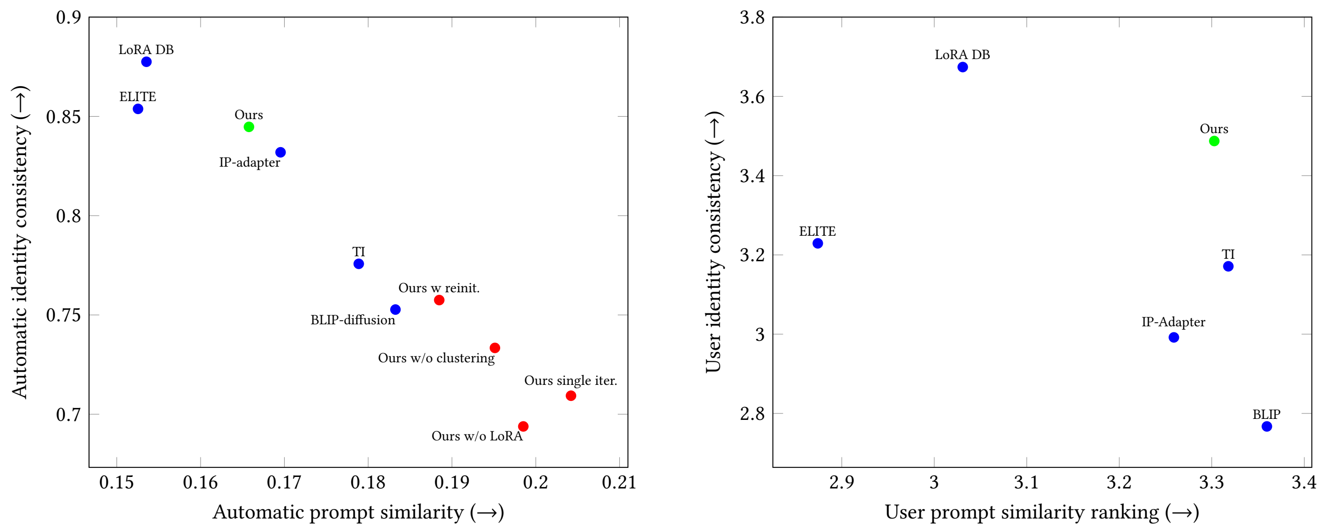


Figure 6: Quantitative Comparison and User Study. (Left) We compared our method quantitatively with various baselines in terms of identity consistency and prompt similarity, as explained in Section 4.1. LoRA DB and ELITE maintain high identity consistency, while sacrificing prompt similarity. TI and BLIP-diffusion achieve high prompt similarity but low identity consistency. We also ablated some components of our method: removing the clustering stage, reducing the optimizable representation, re-initializing the representation in each iteration and performing only a single iteration. All of the ablated cases resulted in a significant degradation of consistency. (Right) The user study rankings also demonstrate that our method is balancing between identity consistency and prompt similarity.

DB. Furthermore, we compared against all available image encoder techniques that encode a single image into the textual space of the diffusion model for later generation in novel contexts: BLIP-Diffusion [Li et al. 2023], ELITE [Wei 2023], and IP-adapter [Ye et al. 2023]. For all the baselines, we used the same prompt p to generate a single image, and used it to extract the identity via optimization (TI and LoRA DB) or encoding (ELITE, BLIP-diffusion and IP-adapter).

In Figure 5 we qualitatively compare our method against the above baselines. While TI [Gal et al. 2022], BLIP-diffusion [Li et al. 2023] and IP-adapter [Ye et al. 2023] are able to follow the specified prompt, they fail to produce a consistent character. LoRA DB [Ryu 2022] succeeds in consistent generation, but it does not always respond to the prompt. Furthermore, the resulting character is generated in the same fixed pose. ELITE [Wei et al. 2023] struggles with prompt following and the generated characters tend to be deformed. In comparison, our method is able to follow the prompt and maintain consistency, while generating appealing characters in different poses and viewing angles.

In order to automatically evaluate our method and the baselines quantitatively, we instructed ChatGPT [OpenAI 2022] to generate prompts for characters of different types (e.g., animals, creatures, objects, etc.) in different styles (e.g., stickers, animations, photorealistic images, etc.). Each of these prompts was then used to extract a consistent character by our method and by each of the baselines. Next, we generated these characters in a predefined collection of novel contexts. For a visual comparison, please refer to the supplementary material.

We employ two standard evaluation metrics: prompt similarity and identity consistency, which are commonly used in the personalization literature [Avrahami et al. 2023a; Gal et al. 2022; Ruiz et al. 2023]. Prompt similarity measures the correspondence between the generated images and the input text prompt. We use the standard CLIP [Radford et al. 2021] similarity, i.e., the normalized cosine similarity between the CLIP image embedding of the generated images and the CLIP text embedding of the source prompts. For measuring identity consistency, we calculate the pairwise similarity between the CLIP image embeddings of generated images of the same concept across different contexts (i.e., when using different text prompts for the same character).

As can be seen in Figure 6 (left), there is an inherent trade-off between prompt similarity and identity consistency: LoRA DB and ELITE exhibit high identity consistency, while sacrificing prompt similarity. TI and BLIP-diffusion achieve high prompt similarity but low identity consistency. Our method achieves better identity consistency than IP-adapter, which is significant from the user's perspective, as supported by our user study.

4.2 User Study

We conducted a user study to evaluate our method, using the Amazon Mechanical Turk (AMT) platform [Amazon 2023]. We used the same generated prompts and samples that were used in Section 4.1 and asked the evaluators to rate the prompt similarity and identity consistency of each result on a Likert scale of 1–5. For ranking the prompt similarity, the evaluators were presented with the target text prompt and the result of our method and the baselines on the same page, and were asked to rate each of the images. For identity consistency, for each of the generated concepts, we compared our method and the baselines by randomly choosing pairs of generated images with different target prompts, and the evaluators were asked to rate on a scale of 1–5 whether the images contain the same main character. Again, all the pairs of the same character for the different baselines were shown on the same page.

As can be seen in Figure 6 (right), our method again exhibits a good balance between identity consistency and prompt similarity, with a wider gap separating it from the baselines. For more details and statistical significance analysis, read the supplementary material.

4.3 Ablation Study

We conducted an ablation study for the following cases: (1) *Without clustering* — we omit the clustering step described in Section 3.1, and instead simply generate 5 images according to the input prompt. (2) *Without LoRA* — we reduce the optimizable representation Θ in the identity extraction stage, as described in Section 3.2, to consist of only the newly-added text tokens without the additional LoRA weights. (3) *With re-initialization* — instead of using the latest representation Θ in each of the optimization iterations, as described in Section 3.3, we re-initialize it in each iteration. (4) *Single iteration* — rather than iterating until convergence (Section 3.3), we stop after a single iteration.

As can be seen in Figure 6 (left), all of the above key components are crucial for achieving a consistent identity in the final result: (1) removing the clustering harms the identity extraction stage because the training set is too diverse, (2) reducing the representation causes underfitting, as the model does not have enough parameters to properly capture the identity, (3) re-initializing the representation in each iteration, or (4) performing a single iteration, does not allow the model to converge into a single identity.

For a visual comparison of the ablation study, as well as comparison of alternative feature extractors (DINOv1 [Caron et al. 2021] and CLIP [Radford et al. 2021]), please refer to the supplementary material.

4.4 Applications

As demonstrated in Figure 7, our method can be used for various down-stream tasks, such as (a) Illustrating a story by breaking it into a different scenes and using the same consistent character for all of them. (b) Local text-driven image editing by integrating Blended Latent Diffusion [Avrahami et al. 2023b, 2022] — a consistent character can be injected into a specified location of a provided background image, in a novel pose specified by a text prompt. (c) Generating a consistent character with an additional pose control using ControlNet [Zhang et al. 2023a]. For more details, please refer to the supplementary material. In addition, as demonstrated in Figure 9 instead of choosing the most cohesive cluster automatically, as explained in Section 3.1, a user can manually select one of the clusters according to their preferences, to affect the final result.

5 LIMITATIONS AND CONCLUSIONS

We found our method to suffer from the following limitations: (a) Inconsistent identity — in some cases, our method is not able to converge to a fully consistent identity (without overfitting). As

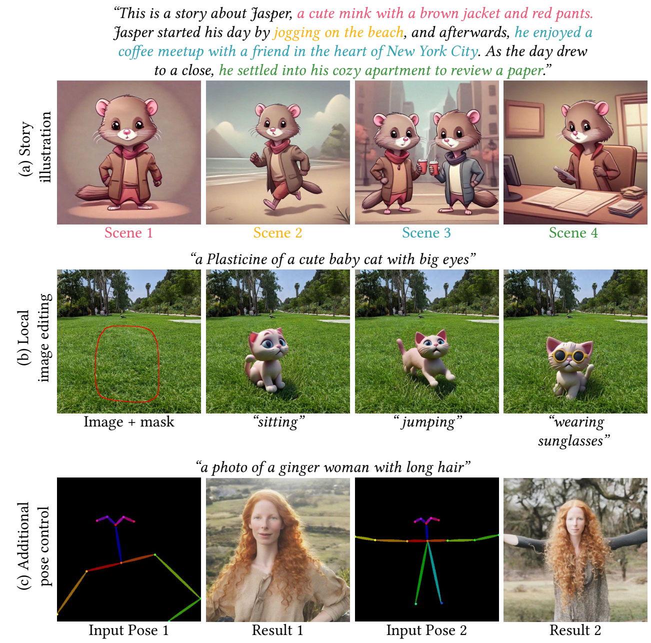


Figure 7: Applications. Our method can be used for various applications: (a) Illustrating a full story with the same consistent character. (b) Local text-driven image editing via integration with Blended Latent Diffusion [Avrahami et al. 2023b, 2022]. (c) Generating a consistent character with an additional pose control via integration with ControlNet [Zhang et al. 2023a].

demonstrated in Figure 8(a), when trying to generate a portrait of a robot, our method generated robots with slightly different colors and shapes (e.g., different arms). This may result from a prompt that is too general, for which identity clustering (Section 3.1) is not able to find a sufficiently cohesive set. (b) Inconsistent supporting characters/elements — although our method is able to find a consistent identity for the character described by the input prompt, the identities of other characters, related to the input character (e.g., their pet), might be inconsistent. For example, in Figure 8(b) the input prompt p to our method described only the girl, and when asked to generate the girl with her cat, different cats were generated. In addition, our framework does not support finding multiple concepts concurrently [Avrahami et al. 2023a]. (c) Spurious attributes — we found that in some cases, our method binds additional attributes, which are not part of the input text prompt, with the final identity of the character. For example, in Figure 8(c), the input text prompt was *"a sticker of a ginger cat"*, however, our method added green leaves to the generated sticker, even though it was not asked to do so. This stems from the stochastic nature of the text-to-image model — the model added these leaves in some of the stickers generated during the identity clustering stage (Section 3.1), and the stickers containing the leaves happened to form the most cohesive set $c_{cohesive}$. One way to mitigate it is to let the user choose one of the most cohesive clusters according to their preferences, instead of selecting it automatically. (d) Significant computational

Figure 8: Limitations. Our method suffers from the following limitations: (a) in some cases, our method is not able to converge to a fully consistent identity — notice slight color and arm shape changes. (b) Our method is not able to associate a consistent identity to a supporting character that may appear with the main extracted character, for example our method generates different cats for the same girl. (c) Our method sometimes adds spurious attributes to the character, that were not present in the text prompt. For example, it learns to associate green leaves with the cat sticker.

cost — each iteration of our method involves generating a large number of images, and learning the identity of the most cohesive cluster. It takes about 20 minutes to converge into a consistent identity. Reducing the computational costs is an appealing direction for further research. (e) Simplistic characters — we found that our method tends to generate simplistic scences (single and mostly centered objects), which may be caused by the "averaging" effect of the identity extraction stage, as explained in Section 3.2.

In conclusion, in this paper we offered the first fully-automated solution to the problem of consistent character generation. We hope that our work will pave the way for future advancements, as we believe this technology of consistent character generation may have a disruptive effect on numerous sectors, including education, storytelling, entertainment, fashion, brand design, advertising, and more.

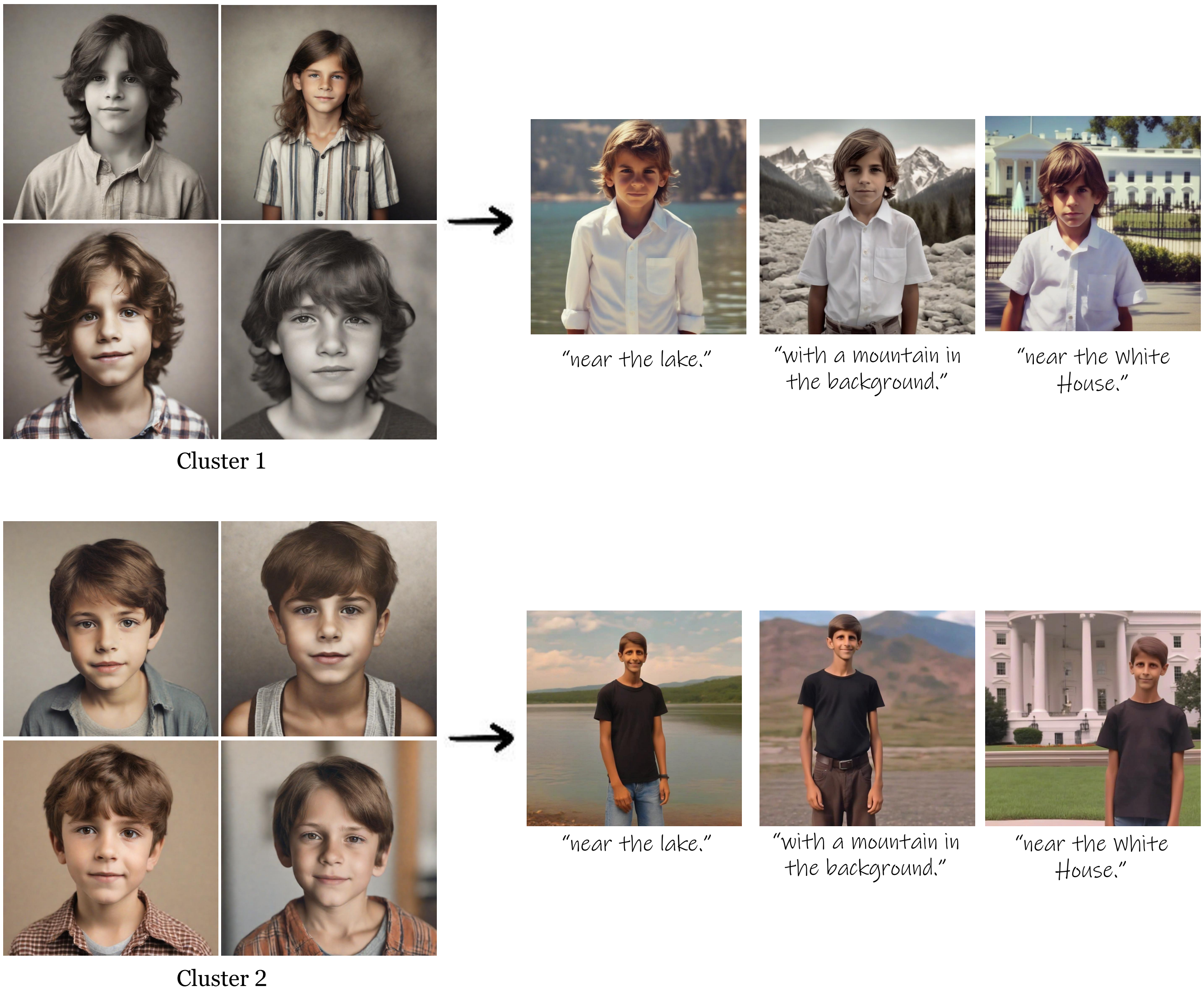


Figure 9: User control. Instead of choosing the most cohesive cluster automatically, as explained in Section 3.1, a user can manually select one of the clusters according to their preferences, to affect the final result. For example, given the text prompt *"a photo of a boy with brown hair"*, the user can control the hairstyle of the generated character by choosing the appropriate cluster.

REFERENCES

Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. 2023. DreamStyler: Paint by Style Inversion with Text-to-Image Diffusion Models. *ArXiv* abs/2309.06933 (2023). https://api.semanticscholar.org/CorpusID:261706081

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. 2023. A Neural Space-Time Representation for Text-to-Image Personalization. *ArXiv* abs/2305.15391 (2023). https://api.semanticscholar.org/CorpusID:258866047

Amazon. 2023. Amazon Mechanical Turk. https://www.mturk.com/.

Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H Bermano. 2023. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06925* (2023).

David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*. https://api.semanticscholar.org/CorpusID:1782131

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023a. Break-A-Scene: Extracting Multiple Concepts from a Single Image. *ArXiv* abs/2305.16311 (2023). https://api.semanticscholar.org/CorpusID:258888228

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023b. Blended Latent Diffusion. *ACM Trans. Graph.* 42, 4, Article 149 (jul 2023), 11 pages. https://doi.org/10.1145/3592450

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023c. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18370–18380.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-Driven Editing of Natural Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18208–18218.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *ArXiv* abs/2211.01324 (2022). https://api.semanticscholar.org/CorpusID:253254800

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*. Springer, 707–723.

Sagie Benaim, Frederik Warburg, Peter Ebert Christensen, and Serge J. Belongie. 2022. Volumetric Disentanglement for 3D Scene Manipulation. *ArXiv* abs/2206.02776 (2022). https://api.semanticscholar.org/CorpusID:249394623

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22560–22570.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9630–9640.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Transactions on Graphics (TOG)* 42 (2023), 1 – 10. https://api.semanticscholar.org/CorpusID:256416326

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. 2023a. Subject-driven Text-to-Image Generation via Apprenticeship Learning. *ArXiv* abs/2304.00186 (2023).

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2023b. AnyDoor: Zero-shot Object-level Image Customization. *ArXiv* abs/2307.09481 (2023). https://api.semanticscholar.org/CorpusID:259951373

Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. *ArXiv* abs/2306.13754 (2023). https://api.semanticscholar.org/CorpusID:259252153

AI Foundations. 2023. How to Create Consistent Characters in Midjourney. https://www.youtube.com/watch?v=Z7_ta3RHijQ.

Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. 2023. SceneScape: Text-Driven Consistent Scene Generation. *ArXiv* abs/2302.01133 (2023). https://api.semanticscholar.org/CorpusID:256503775

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–13.

Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. 2023. Expressive Text-to-Image Generation with Rich Text. *ArXiv* abs/2304.06720 (2023). https://api.semanticscholar.org/CorpusID:258108187

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373* (2023).

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. TaleCrafter: Interactive Story Visualization with Multiple Characters. *ArXiv* abs/2305.18247 (2023). https://api.semanticscholar.org/CorpusID:258960665

Ori Gordon, Omri Avrahami, and Dani Lischinski. 2023. Blended-NeRF: Zero-Shot Object Generation and Blending in Existing Neural Radiance Fields. *ArXiv* abs/2306.12760 (2023). https://api.semanticscholar.org/CorpusID:259224726

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris N. Metaxas, and Feng Yang. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. *ArXiv* abs/2303.11305 (2023).

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. 2023. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2328–2337.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

Geoffrey E. Hinton and Sam T. Roweis. 2002. Stochastic Neighbor Embedding. In *NIPS*. https://api.semanticscholar.org/CorpusID:20240

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Proc. NeurIPS*.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. 2023. Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. *ArXiv* abs/2303.11989 (2023). https://api.semanticscholar.org/CorpusID:257636653

Eliahu Horwitz and Yedid Hoshen. 2022. Conffusion: Confidence Intervals for Diffusion Models. *ArXiv* abs/2211.09795 (2022).

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP. https://doi.org/10.5281/zenodo.5143773

Shira Iluz, Yael Vinker, Amir Hertz, Daniel Berio, Daniel Cohen-Or, and Ariel Shamir. 2023. Word-As-Image for Semantic Typography. *ACM Transactions on Graphics (TOG)* 42 (2023), 1 – 11. https://api.semanticscholar.org/CorpusID:257353586

Hyeonho Jeong, Gihyun Kwon, and Jong-Chul Ye. 2023. Zero-shot Generation of Coherent Storybook from Plain Text Story using Diffusion Models. *ArXiv* abs/2302.03900 (2023). https://api.semanticscholar.org/CorpusID:256662241

Xuhui Jia, Yang Zhao, Kelvin C. K. Chan, Yandong Li, Han-Ying Zhang, Boqing Gong, Tingbo Hou, H. Wang, and Yu-Chuan Su. 2023. Taming Encoder for Zero Fine-tuning Image Customization with Text-to-Image Diffusion Models. *ArXiv* abs/2304.02642 (2023).

JoshGreat. 2023. 8 ways to generate consistent characters (for comics, storyboards, books etc) : StableDiffusion. https://www.reddit.com/r/StableDiffusion/comments/10yxz3m/8_ways_to_generate_consistent_characters_for/.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

William H. Kruskal and Wilson Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47 (1952), 583–621.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

Dongxu Li, Junnan Li, and Steven C. H. Hoi. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. *ArXiv* abs/2305.14720 (2023). https://api.semanticscholar.org/CorpusID:258865473

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. StoryGAN: A Sequential Conditional GAN for Story Visualization. *CVPR* (2019).

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2023a. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761* (2023).

Shaoteng Liu, Yuecheng Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2023b. Video-P2P: Video Editing with Cross-attention Control. *ArXiv* abs/2303.04761 (2023). https://api.semanticscholar.org/CorpusID:257405406

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pre-trained text-to-image transformers for story continuation. In *European Conference on Computer Vision*. Springer, 70–87.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic

Differential Equations. In *International Conference on Learning Representations.*

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12663–12673.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6038–6047.

Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Y. Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. 2023. Dreamix: Video Diffusion Models are General Video Editors. *ArXiv* abs/2302.01329 (2023).

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453* (2023).

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning.* https://api.semanticscholar.org/CorpusID:245335086

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision. *ArXiv* abs/2304.07193 (2023). https://api.semanticscholar.org/CorpusID:258170077

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Localizing Object-level Shape Variations with Text-to-Image Diffusion Models. *ArXiv* abs/2303.11306 (2023).

Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T. Barron, Amit H. Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, C. Karen Liu, Lingjie Liu, Ben Mildenhall, Matthias Nießner, Bjorn Ommer, Christian Theobalt, Peter Wonka, and Gordon Wetzstein. 2023. State of the Art on Diffusion Models for Visual Computing. *ArXiv* abs/2310.07204 (2023). https://api.semanticscholar.org/CorpusID:263835355

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *ArXiv* abs/2307.01952 (2023). https://api.semanticscholar.org/CorpusID:259341735

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535* (2023).

Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H. Bermano, and Daniel Cohen-Or. 2023. Single Motion Diffusion. *ArXiv* abs/2302.05905 (2023). https://api.semanticscholar.org/CorpusID:256827051

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning.*

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, S. Tulyakov, Shweta Mahajan, and Leonid Sigal. 2022. Make-A-Story: Visual Memory Conditioned Consistent Story Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 2493–2502. https://api.semanticscholar.org/CorpusID:254017562

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125* (2022).

Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. 2023a. ConceptLab: Creative Generation using Diffusion Prior Constraints. *arXiv preprint arXiv:2308.02669* (2023).

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023b. TEXTure: Text-Guided Texturing of 3D Shapes. *ACM SIGGRAPH 2023 Conference Proceedings* (2023). https://api.semanticscholar.org/CorpusID:256597953

Romain Beaumont 2023. CLIP Retrival. https://github.com/rom1504/clip-retrieval.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10674–10685.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 22500–22510.

Simo Ryu. 2022. Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning. https://github.com/cloneofsimo/lora.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *ArXiv* abs/2210.08402 (2022). https://api.semanticscholar.org/CorpusID:252917726

Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. 2023. Vox-E: Text-guided Voxel Editing of 3D Objects. *ArXiv* abs/2303.12048 (2023). https://api.semanticscholar.org/CorpusID:257636627

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2022. kNN-Diffusion: Image Generation via Large-Scale Retrieval. In *The Eleventh International Conference on Learning Representations.*

Jing Shi, Wei Xiong, Zhe L. Lin, and Hyun Joon Jung. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. *ArXiv* abs/2304.03411 (2023).

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning.* PMLR, 2256–2265.

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. 2023. StyleDrop: Text-to-Image Generation in Any Style. *ArXiv* abs/2306.00983 (2023). https://api.semanticscholar.org/CorpusID:258999204

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations.*

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems* 32 (2019).

stassius. 2023. How to create consistent character faces without training (info in the comments) : StableDiffusion. https://www.reddit.com/r/StableDiffusion/comments/12djxvz/how_to_create_consistent_character_faces_without/.

Gábor Szűcs and Modafar Al-Shouha. 2022. Modular StoryGAN with background and theme awareness for story visualization. In *International Conference on Pattern Recognition and Artificial Intelligence.* Springer, 275–286.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. *ArXiv* abs/2209.14916 (2022). https://api.semanticscholar.org/CorpusID:252595883

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. *ACM SIGGRAPH 2023 Conference Proceedings* (2023). https://api.semanticscholar.org/CorpusID:258436985

John W. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* 5 2 (1949), 99–114.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1921–1930.

Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. 2023. Face0: Instantaneously Conditioning a Text-to-Image Model on a Face. *SIGGRAPH Asia 2023 Conference Papers* (2023). https://api.semanticscholar.org/CorpusID:259138505

Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept Decomposition for Visual Exploration and Inspiration. *ArXiv* abs/2305.18203 (2023). https://api.semanticscholar.org/CorpusID:258959472

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2022. Sketch-Guided Text-to-Image Diffusion Models. *arXiv preprint arXiv:2211.13752* (2022).

Andrey Voynov, Q. Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. P+: Extended Textual Conditioning in Text-to-Image Generation. *ArXiv* abs/2303.09522 (2023).

Yuxiang Wei. 2023. Official Implementation of ELITE. https://github.com/csyxwei/ELITE. Accessed: 2023-05-01.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. *ArXiv* abs/2302.13848 (2023).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. *ArXiv* abs/2306.07954 (2023). https://api.semanticscholar.org/CorpusID:259144797

Hu Ye, Jun Zhang, Siyi Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *ArXiv* abs/2308.06721

(2023). https://api.semanticscholar.org/CorpusID:260886966

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *arXiv preprint arXiv:2206.10789* (2022).

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In-So Kweon. 2023b. Text-to-image Diffusion Models in Generative AI: A Survey. *ArXiv* abs/2303.07909 (2023). https://api.semanticscholar.org/CorpusID:257505012

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3836–3847.

Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. 2023. DreamEditor: Text-Driven 3D Scene Editing with Neural Fields. *ArXiv* abs/2306.13455 (2023). https://api.semanticscholar.org/CorpusID:259243782

ACKNOWLEDGMENTS

We thank Yael Pitch, Matan Cohen, Neal Wadhwa and Yaron Brodsky for their valuable help and feedback.

A ADDITIONAL EXPERIMENTS

Below, we provide additional experiments that were omitted from the main paper. In Appendix A.1 we provide additional comparisons and results of our method, and demonstrate its non-deterministic nature in Appendix A.2. In Appendix A.3 we compare our method against two naïve baselines. Appendix A.4 presents the results of our method using different feature extractors. Lastly, in Appendix A.6 we provide results that reduce the concerns of dataset memorization by our method.

A.1 Additional Comparisons and Results

In Figure 10 we provide a qualitative comparison on the automatically generated prompts, and in Figure 11 we provide an additional qualitative comparison.

Concurrently to our work, the DALL·E 3 model [Betker et al. 2023] was commercially released as part of the paid ChatGPT Plus [OpenAI 2022] subscription, enabling generating images in a conversational setting. We tried, using a conversation, to create a consistent character of a Plasticine cat, as demonstrated in Figure 20. As can be seen, the generated characters share only some of the characteristics (e.g., big eyes) but not all of them (e.g., colors, textures and shapes).

In Figure 12 we provide a qualitative comparison of the ablated cases. In addition, as demonstrated in Figure 13, our approach is applicable to consistent generation of a wide range of subjects, without the requirement for them to necessarily depict human characters or creatures. Figure 14 shows additional results of our method, demonstrating a variety of character styles. Lastly, in Figure 15 we demonstrate the ability of creating a fully consistent "life story" of a character using our method.

A.2 Non-determinism of Our Method

In Figures 16 and 17 we demonstrate the non-deterministic nature of our method. Using the same text prompt, we run our method multiple times with different initial seeds, thereby generating a different set of images for the identity clustering stage (Section 3.1). Consequently, the most cohesive cluster $c_{cohesive}$ is different in each run, yielding different consistent identities. This behavior of our method is aligned with the one-to-many nature of our task — a single text prompt may correspond to many identities.

A.3 Naïve Baselines

As explained in Section 4.1, we compared our method against a version of TI [Gal et al. 2022] and LoRA DB [Ryu 2022] that were trained on a single image (with a single identity). Instead, we could generate a small set of five images for the given prompt (that are not guaranteed to be of the same identity), and use this small dataset for TI and LoRA DB baselines, referred to as *TI multi* and *LoRA DB multi*, respectively. As can be seen in Figures 18 and 19, these baselines fail to achieve satisfactory identity consistency.

A.4 Additional Feature Extractors

Instead of using DINOv2 [Oquab et al. 2023] features for the identity clustering stage (Section 3.1), we also experimented with two alternative feature extractors: DINOv1 [Caron et al. 2021] and CLIP [Radford et al. 2021] image encoder. We quantitatively evaluate our method with each of these feature extractors in terms of identity consistency and prompt similarity, as explained in Section 4.1. As can be seen in Figure 21, DINOv1 produces higher identity consistency, while sacrificing prompt similarity, whereas CLIP achieves higher prompt similarity at the expense of identity consistency. Qualitatively, as demonstrated in Figure 22, we found the DINOv1 extractor to perform similarly to DINOv2, whereas CLIP produces results with a slightly lower identity consistency.

A.5 Additional Clustering Visualization

In Figure 23 we provide a visualization of the clustering algorithm described in Section 3.1. As can be seen, given the input text prompt *"a purple astronaut, digital art, smooth, sharp focus, vector art"*, in the first iteration (top three rows), our algorithm divides the generated image set into three clusters: (1) focusing on the astronaut's head, (2) an astronaut with no face, and (3) a full body astronaut. In the second iteration (bottom three rows), all the clusters share the same identity, that was extracted in the first iteration, as described in Section 3.2, and our algorithm divides them into clusters by their pose.

A.6 Dataset Non-Memorization

Our method is able to produce consistent characters, which raises the question of whether these characters already exist in the training data of the generative model. We employed SDXL [Podell et al. 2023] as our text-to-image model, whose training dataset is, unfortunately, undisclosed in the paper [Podell et al. 2023]. Consequently, we relied on the most likely overlapping dataset, LAION-5B [Schuhmann et al. 2022], which was also utilized by Stable Diffusion V2.

To probe for dataset memorization, we found the top 5 nearest neighbors in the dataset in terms of CLIP [Radford et al. 2021] image similarity, for a few representative characters from our paper, using an open-source solution [Romain Beaumont 2023]. As demonstrated in Figure 24, our method does not simply memorize images from the LAION-5B dataset.

A.7 Stable Diffusion 2 Results

We experimented with a version of our method that uses the Stable Diffusion 2 [Rombach et al. 2021] model. The implementation is the same as explained in Appendix B.1, with the following changes: (1) The set of custom text embeddings τ in the character representation

Θ (as explained in Section 2 in the main paper), contains only one text embedding. (2) We used a higher learning rate of 5e-4. The rest of the implementation details are the same. More specifically, we used Stable Diffusion v2.1 implementation from Diffusers [von Platen et al. 2022] library.

As can be seen in Figure 25, when using the Stable Diffusion 2 backbone, our method can extract a consistent character, however, as expected, the results are of a lower quality than when using the SDXL [Podell et al. 2023] backbone that we use in the rest of this paper.

B IMPLEMENTATION DETAILS

In this section, we provide the implementation details that were omitted from the main paper. In Appendix B.1 we provide the implementation details of our method and the baselines. Then, in Appendix B.2 we provide the implementation details of the automatic metrics that we used to evaluate our method against the baselines. In Appendix B.3 we provide the implementation details and the statistical analysis for the user study we conducted. Lastly, in Appendix B.4 we provide the implementation details for the applications we presented.

B.1 Method Implementation Details

We based our method, and all the baselines (except ELITE [Wei et al. 2023] and BLIP-diffusion [Li et al. 2023]) on Stable Diffusion XL (SDXL) [Podell et al. 2023], which is the state-of-the-art open source text-to-image model, at the writing of this paper. We used the official ELITE implementation, that uses Stable Diffusion V1.4, and the official implementation of BLIP-diffusion, that uses Stable Diffusion V1.5. We could not replace these two baselines to SDXL backbone, as the encoders were trained on these specific models. As for the rest of the baselines, we used the same SDXL architecture and weights.

For our method, we generated a set of $N = 128$ images at each iteration, which we found to be sufficient, empirically. We utilized the Adam optimizer [Kingma and Ba 2014] with learning rate of 3e-5, $\beta_1 = 0.9$, $\beta_2 = 0.99$ and weight decay of 1e-2. In each identity extraction iteration of our method, we used 500 steps. We also found empirically that we can set the convergence criterion d_{conv} adaptively to be 80% of the average pairwise Euclidean distance between all N initial image embeddings of the first iteration. In most cases, our method converges in 1–2 iterations, which takes about 13–26 minutes on A100 NVIDIA GPU when using bfloat16 mixed precision. In addition, we found that encouraging small clusters is beneficial by setting the minimum cluster size $d_{min\text{-}c}$, and the target cluster size $d_{size\text{-}c}$ to $d_{min\text{-}c} = d_{size\text{-}c} = 5$, which is the recommended image set size in the personalization setting [Gal et al. 2022; Ruiz et al. 2023].

List of the third-party packages that we used:

- Official SDXL [Podell et al. 2023] implementation by HuggingFace Diffusers [von Platen et al. 2022] at https://github.com/huggingface/diffusers
- Official SDXL LoRA DB implementation by HuggingFace Diffusers [von Platen et al. 2022] at https://github.com/huggingface/diffusers.
- Official ELITE [Wei et al. 2023] implementation at https://github.com/csyxwei/ELITE
- Official BLIP-diffusion [Li et al. 2023] implementation at https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion
- Official IP-adapter [Ye et al. 2023] implementation at https://github.com/tencent-ailab/IP-Adapter
- DINOv2 [Oquab et al. 2023] ViT-g/14, DINOv1 [Caron et al. 2021] ViT-B/16 and CLIP [Radford et al. 2021] ViT-L/14 implementation by HuggingFace Transformers [Wolf et al. 2020] at https://github.com/huggingface/transformers

Table 1: Users' rankings means and variances. The means and variances of the rankings that are reported in the user study.

Method	Prompt similarity (↑)	Identity consistency (↑)
TI [Gal et al. 2022]	3.31 ± 1.43	3.17 ± 1.17
LoRA DB [Ryu 2022]	3.03 ± 1.43	3.67 ± 1.20
ELITE [Wei et al. 2023]	2.87 ± 1.46	3.20 ± 1.21
BLIP-Diffusion [Li et al. 2023]	3.35 ± 1.41	2.76 ± 1.31
IP-Adapter [Ye et al. 2023]	3.25 ± 1.42	2.99 ± 1.28
Ours	3.30 ± 1.36	3.48 ± 1.20

B.2 Automatic Metrics Implementation Details

In order to automatically evaluate our method and the baselines quantitatively, we instructed ChatGPT [OpenAI 2022] to generate prompts for characters of different types (e.g., animals, creatures, objects, etc.) in different styles (e.g., stickers, animations, photorealistic images, etc.). These prompts were then used to generate a set of consistent characters by our method and by each of the baselines. Next, these prompts were used to generate these characters in a predefined collection of novel contexts from the following list:

- "a photo of [v] at the beach"
- "a photo of [v] in the jungle"
- "a photo of [v] in the snow"
- "a photo of [v] in the street"
- "a photo of [v] with a city in the background"
- "a photo of [v] with a mountain in the background"
- "a photo of [v] with the Eiffel Tower in the background"
- "a photo of [v] near the Statue of Liberty"
- "a photo of [v] near the Sydney Opera House"
- "a photo of [v] floating on top of water"
- "a photo of [v] eating a burger"
- "a photo of [v] drinking a beer"
- "a photo of [v] wearing a blue hat"
- "a photo of [v] wearing sunglasses"
- "a photo of [v] playing with a ball"
- "a photo of [v] as a police officer"

where [v] is the newly-added token that represents the consistent character.

B.3 User Study Details

As explained in Section 4.2, we conducted a user study to evaluate our method, using the Amazon Mechanical Turk (AMT) platform [Amazon 2023]. We used the same generated prompts and samples

Table 2: Statistical analysis. We use Tukey's honestly significant difference procedure [Tukey 1949] to test whether the differences between mean scores in our user study are statistically significant.

Method 1	Method 2	Prompt similarity p-value	Identities similarity p-value
TI [Gal et al. 2022]	Ours	$p < 0.001$	$p < 1e{-}10$
LoRA DB [Ryu 2022]	Ours	$p < 1e{-}13$	1e−4
ELITE [Wei et al. 2023]	Ours	$p < 1e{-}13$	$p < 1e{-}7$
BLIP-Diffusion [Li et al. 2023]	Ours	$p < 0.01$	$p < 1e{-}13$
IP-Adapter [Ye et al. 2023]	Ours	$p < 1e{-}5$	$p < 1e{-}13$

that were used in Section 4.1, and asked the evaluators to rate the prompt similarity and identity consistency of each result on a Likert scale of 1–5. For ranking the prompt similarity, the evaluators were instructed the following: "For each of the following images, please rank on a scale of 1 to 5 its correspondence to this text description: {PROMPT}. The character in the image can be anything (e.g., a person, an animal, a toy etc." where {PROMPT} is the target text prompt (in which we replaced the special token with the word "character"). All the baselines, as well as our method, were presented in the same page, and the evaluators were asked to rate each one of the results using a slider from 1 ("Do not match at all") to 5 ("Match perfectly"). Next, to assess identity consistency, we took for each one of the characters two generated images that correspond to *different* target text prompts, put them next to each other, and instructed the evaluators the following: "For each of the following image pairs, please rank on a scale of 1 to 5 if they contain the same character (1 means that they contain totally different characters and 5 means that they contain exactly the same character). The images can have different backgrounds". We put all the compared images on the same page, and the evaluators were asked to rate each one of the pairs using a slider from 1 ("Totally different characters") to 5 ("Exactly the same character").

We collected three ratings per question, resulting in 1104 ratings per task (prompt similarity and identity consistency). The time allotted per task was one hour, to allow the raters to properly evaluate the results without time pressure. The means and variances of the user study responses are reported in Table 1.

In addition, we conducted a statistical analysis of our user study by validating that the difference between all the conditions is statistically significant using Kruskal-Wallis [Kruskal and Wallis 1952] test ($p < 1e{-}28$ for the text similarity test and $p < 1e{-}76$ for the identity consistency text). Lastly, we used Tukey's honestly significant difference procedure [Tukey 1949] to show that the comparison of our method against all the baselines is statistically significant, as detailed in Table 2.

B.4 Applications Implementation Details

In Section 4.4, we presented three downstream applications of our method.

Story illustration. Given a long story, e.g., *"This is a story about Jasper, a cute mink with a brown jacket and red pants. Jasper started his day by jogging on the beach, and afterwards, he enjoyed a coffee meetup with a friend in the heart of New York City. As the day drew to a close, he settled into his cozy apartment to review a paper"*, one can create a consistent character from the main character description (*"a cute mink with a brown jacket and red pants"*), then they can generate the various scenes by simply rephrasing the sentence:

(1) *"[v] jogging on the beach"*
(2) *"[v] drinking coffee with his friend in the heart of New York City"*
(3) *"[v] reviewing a paper in his cozy apartment"*

Local image editing. Our method can be simply integrated with Blended Latent Diffusion [Avrahami et al. 2023b, 2022] for editing images locally: given a text prompt, we start by running our method to extract a consistent identity, then, given an input image and mask, we can plant the character in the image within the mask boundaries. In addition, we can provide a local text description for the character.

Additional pose control. Our method can be integrated with ControlNet [Zhang et al. 2023a]: given a text prompt, we first apply our method to extract a consistent identity $\Theta = (\theta, \tau)$, where θ are the LoRA weights and τ is a set of custom text embeddings. Then, we can take an off-the-shelf pre-trained ControlNet model, plug-in our representation Θ, and use it to generate the consistent character in different poses given by the user.

C SOCIETAL IMPACT

We believe that the emergence of technology that facilitates the effortless creation of consistent characters holds exciting promise in a variety of creative and practical applications. It can empower storytellers and content creators to bring their narratives to life with vivid and unique characters, enhancing the immersive quality of their work. In addition, it may offer accessibility to those who may not possess traditional artistic skills, democratizing character design in the creative industry. Furthermore, it can reduce the cost of advertising, and open up new opportunities for small and underprivileged entrepreneurs, enabling them to reach a wider audience and compete in the market more effectively.

On the other hand, as any other generative AI technology, it can be misused by creating false and misleading visual content for deceptive purposes. Creating fake characters or personas can be used for online scams, disinformation campaigns, etc., making it challenging to discern genuine information from fabricated content. Such technologies underscore the vital importance of developing generated content detection systems, making it a compelling research direction to address. In addition, since our method uses a clustering algorithm, there exists a risk of automatically choosing a cluster with improper content, which may result in creating an improper consistent character.

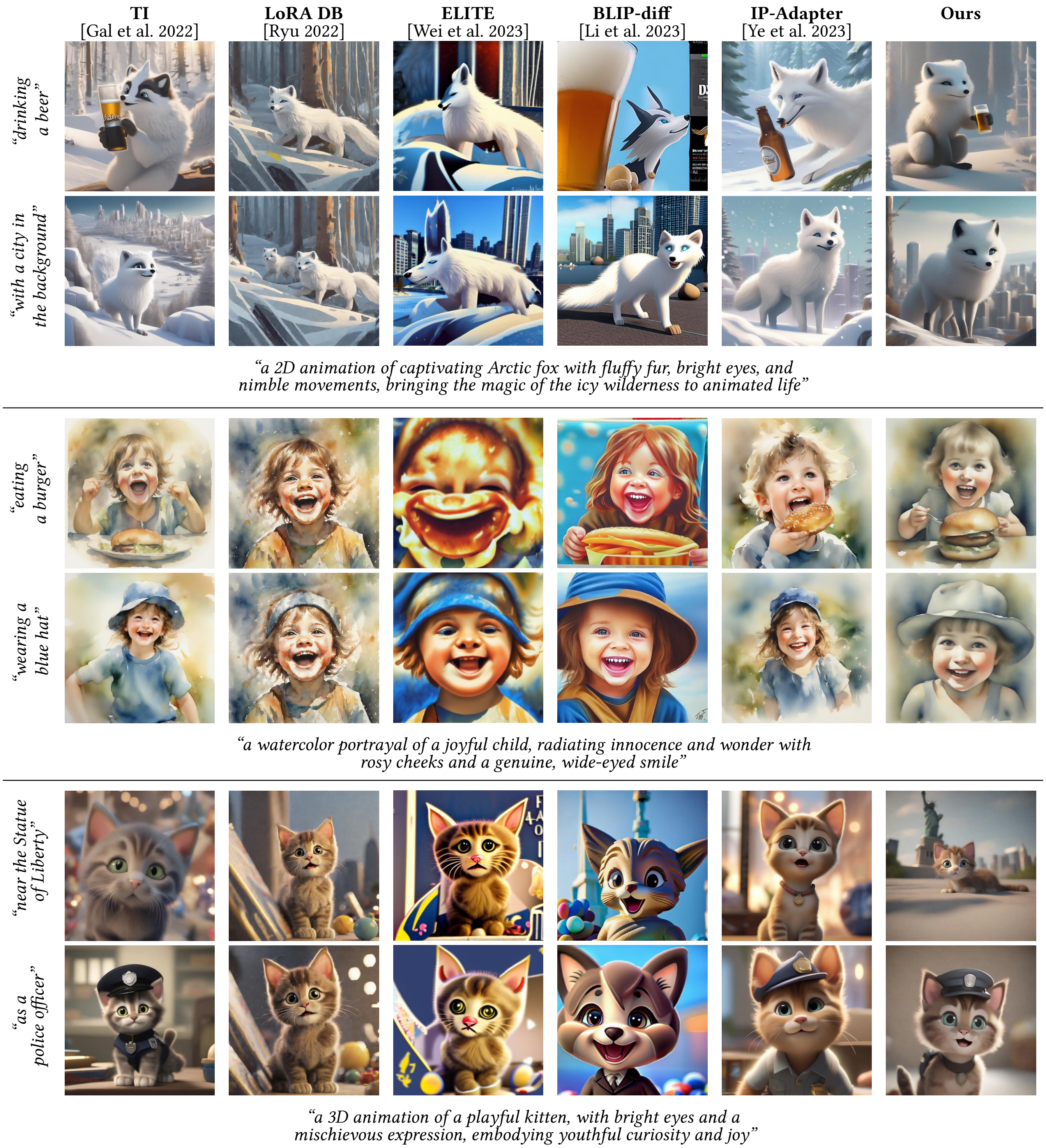


Figure 10: Qualitative comparison to baselines on the automatically generated prompts. We compared our method against several baselines: TI [Gal et al. 2022], BLIP-diffusion [Li et al. 2023] and IP-adapter [Ye et al. 2023] are able to correspond to the target prompt but fail to produce consistent results. LoRA DB [Ryu 2022] is able to achieve consistency, but it does not always follow to the prompt, in addition, the generate character is being generated in the same fixed pose. ELITE [Wei et al. 2023] struggles with following the prompt and also tends to generate deformed characters. Our method is able to follow the prompt, and generate consistent characters in different poses and viewing angles.

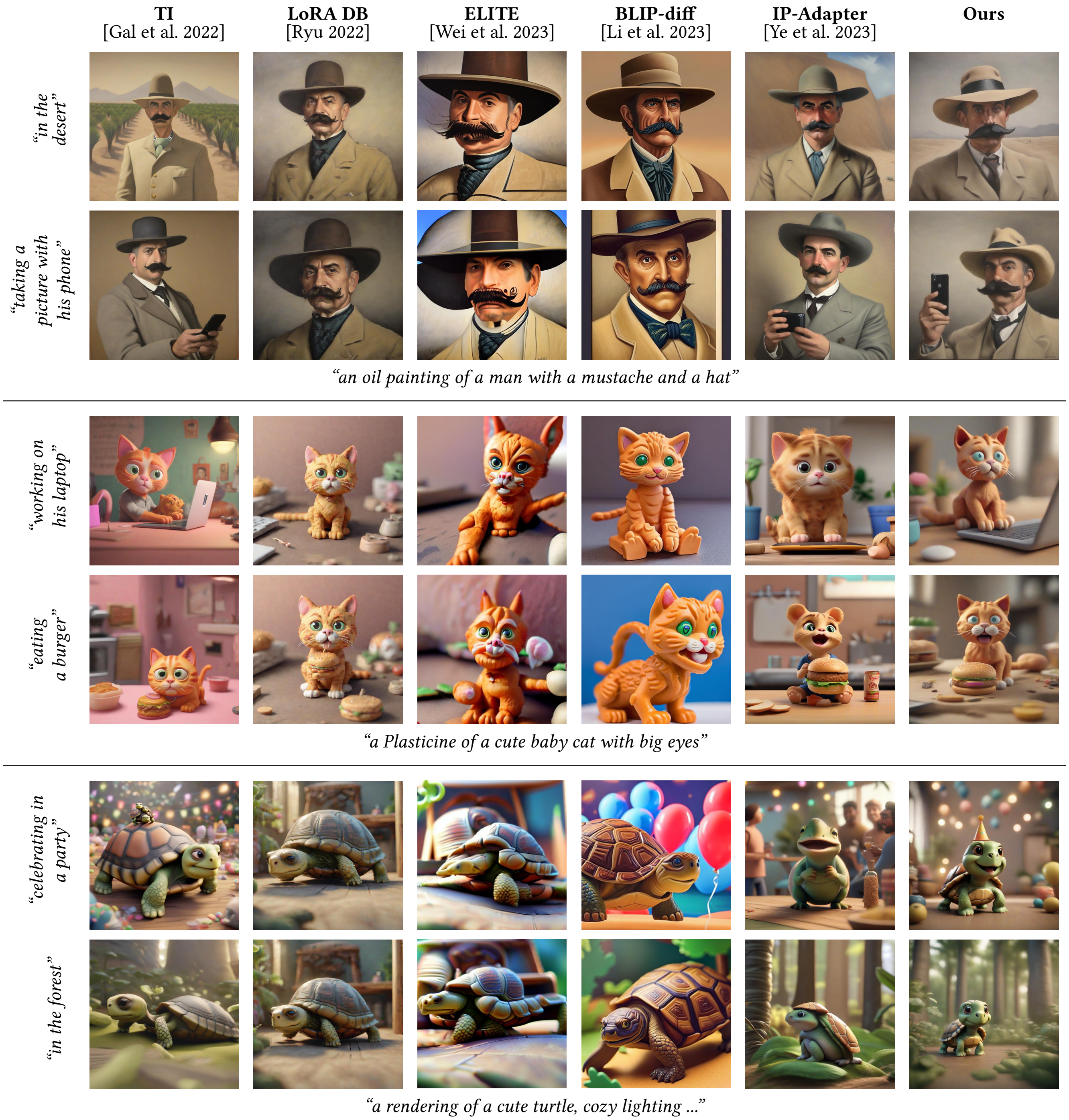


Figure 11: Additional qualitative comparisons to baselines. We compared our method against several baselines: TI [Gal et al. 2022], BLIP-diffusion [Li et al. 2023] and IP-adapter [Ye et al. 2023] are able to correspond to the target prompt but fail to produce consistent results. LoRA DB [Ryu 2022] is able to achieve consistency, but it does not always follow to the prompt, in addition, the generate character is being generated in the same fixed pose. ELITE [Wei et al. 2023] struggles with following the prompt and also tends to generate deformed characters. On the other hand, our method is able to follow the prompt, and generate consistent characters in different poses and viewing angles.

Figure 12: Qualitative comparison of ablations. We ablated the following components of our method: using a single iteration, removing the clustering stage, removing the LoRA trainable parameters, using the same initial representation at every iteration. As can be seen, all these ablated cases struggle with preserving the character's consistency.

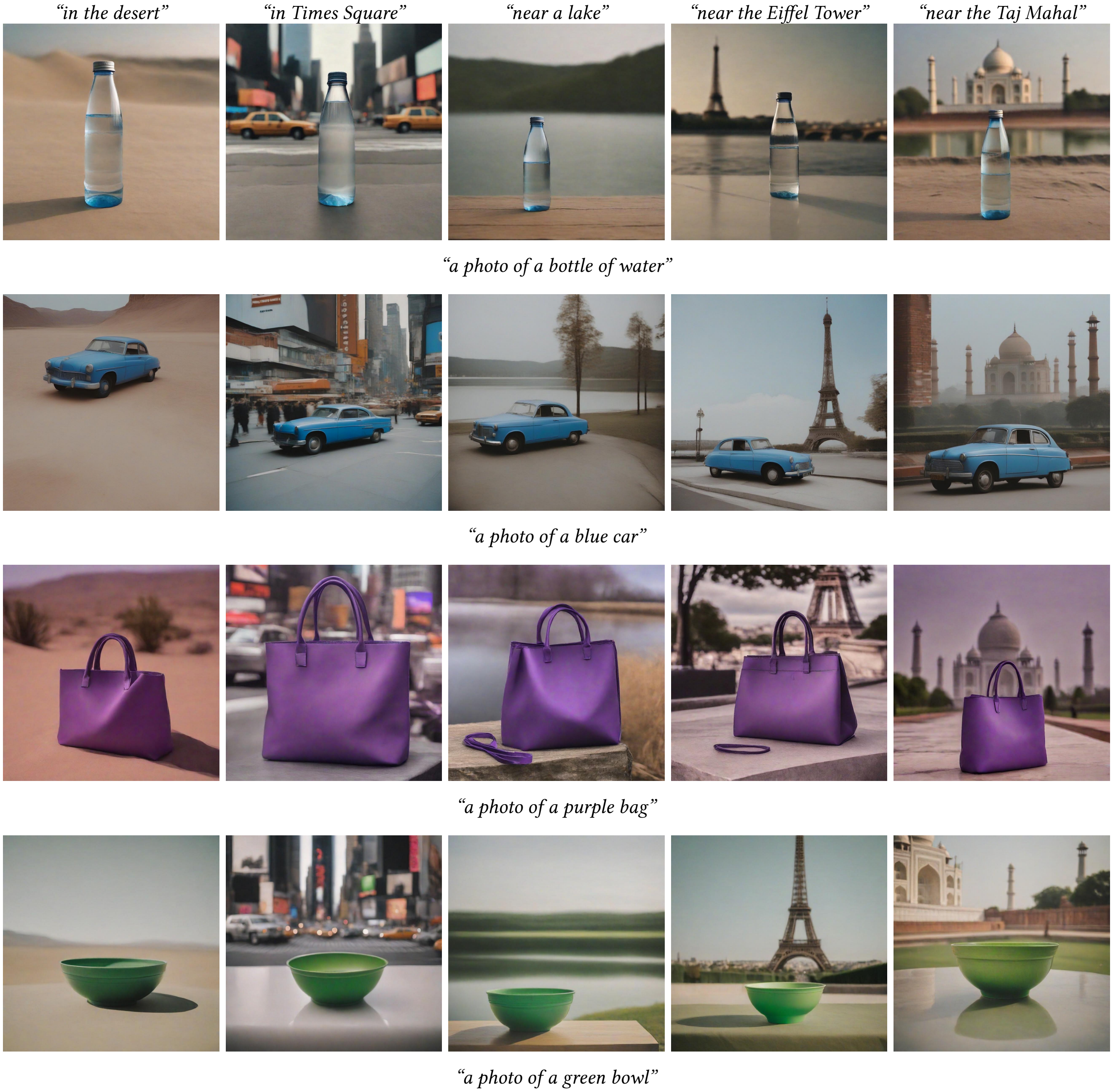
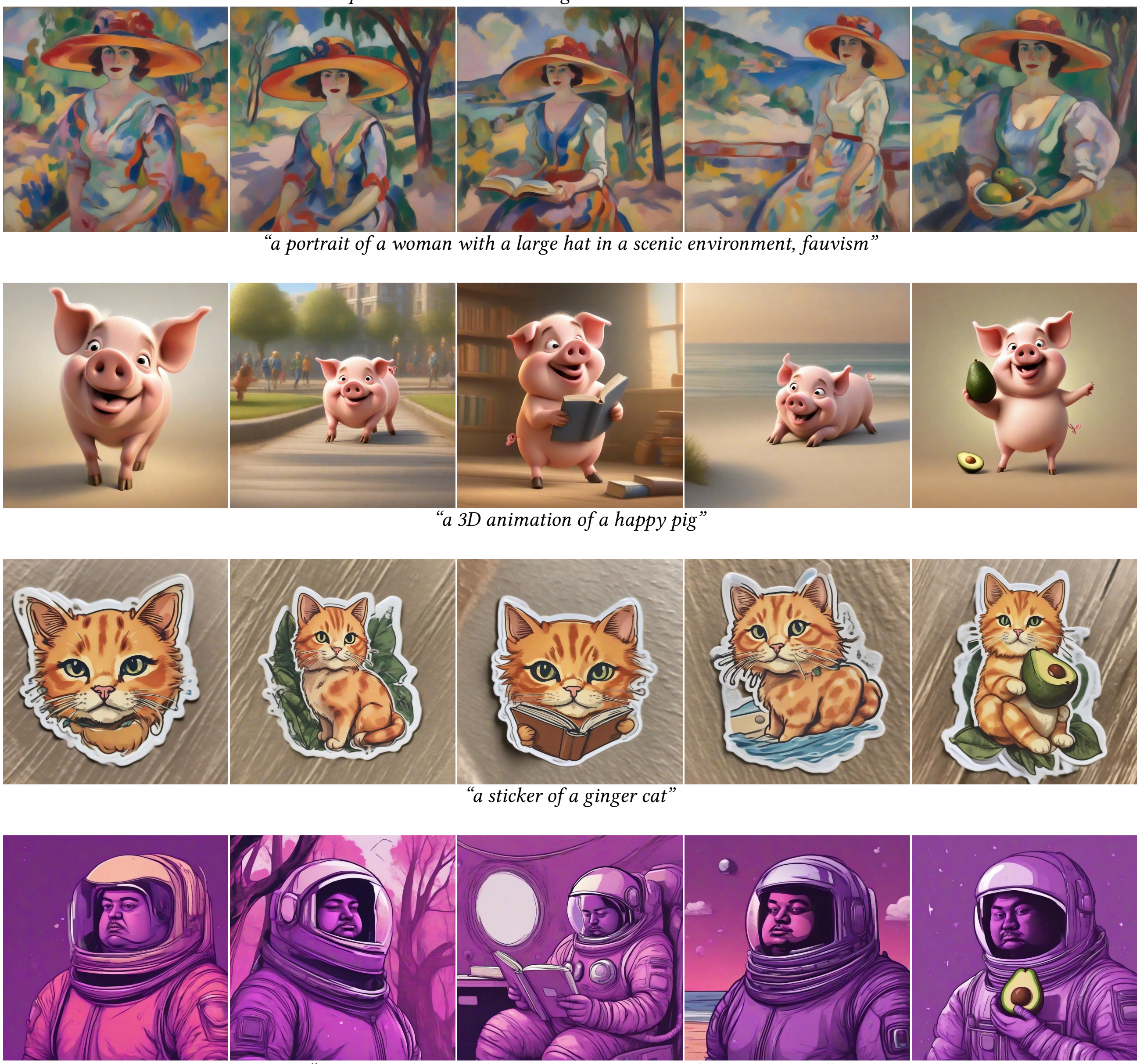


Figure 13: Consistent generation of non-character objects. Our approach is applicable to a wide range of objects, without the requirement for them to depict human characters or creatures.

"in the park" *"reading a book"* *"at the beach"* *"holding an avocado"*



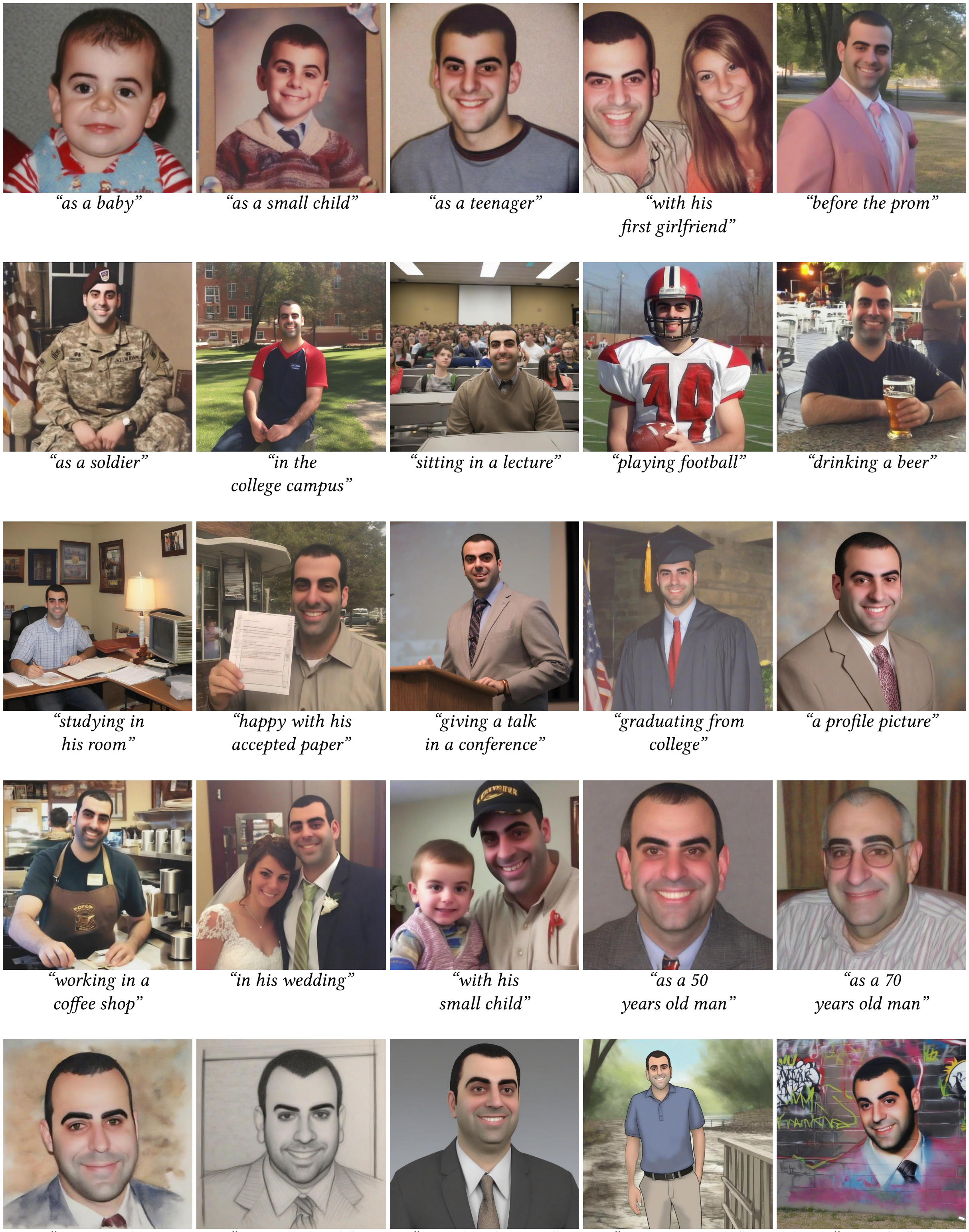
"a portrait of a woman with a large hat in a scenic environment, fauvism"

"a 3D animation of a happy pig"

"a sticker of a ginger cat"

"a purple astronaut, digital art, smooth, sharp focus, vector art"

Figure 14: Additional results. Our method is able to consistently generate different types and styles of characters, e.g., paintings, animations, stickers and vector art.



"as a baby" *"as a small child"* *"as a teenager"* *"with his first girlfriend"* *"before the prom"*

"as a soldier" *"in the college campus"* *"sitting in a lecture"* *"playing football"* *"drinking a beer"*

"studying in his room" *"happy with his accepted paper"* *"giving a talk in a conference"* *"graduating from college"* *"a profile picture"*

"working in a coffee shop" *"in his wedding"* *"with his small child"* *"as a 50 years old man"* *"as a 70 years old man"*

"a watercolor painting" *"a pencil sketch"* *"a rendered avatar"* *"a 2D animation"* *"a graffiti"*

Figure 15: Life story. Given a text prompt describing a fictional character, *"a photo of a man with short black hair"*, we can generate a consistent life story for that character, demonstrating the applicability of our method for story generation.

Figure 16: Non-determinism. By running our method multiple times, given the same prompt *"a photo of a 50 years old man with curly hair"*, but using different initial seeds, we obtain different consistent characters corresponding to the text prompt.

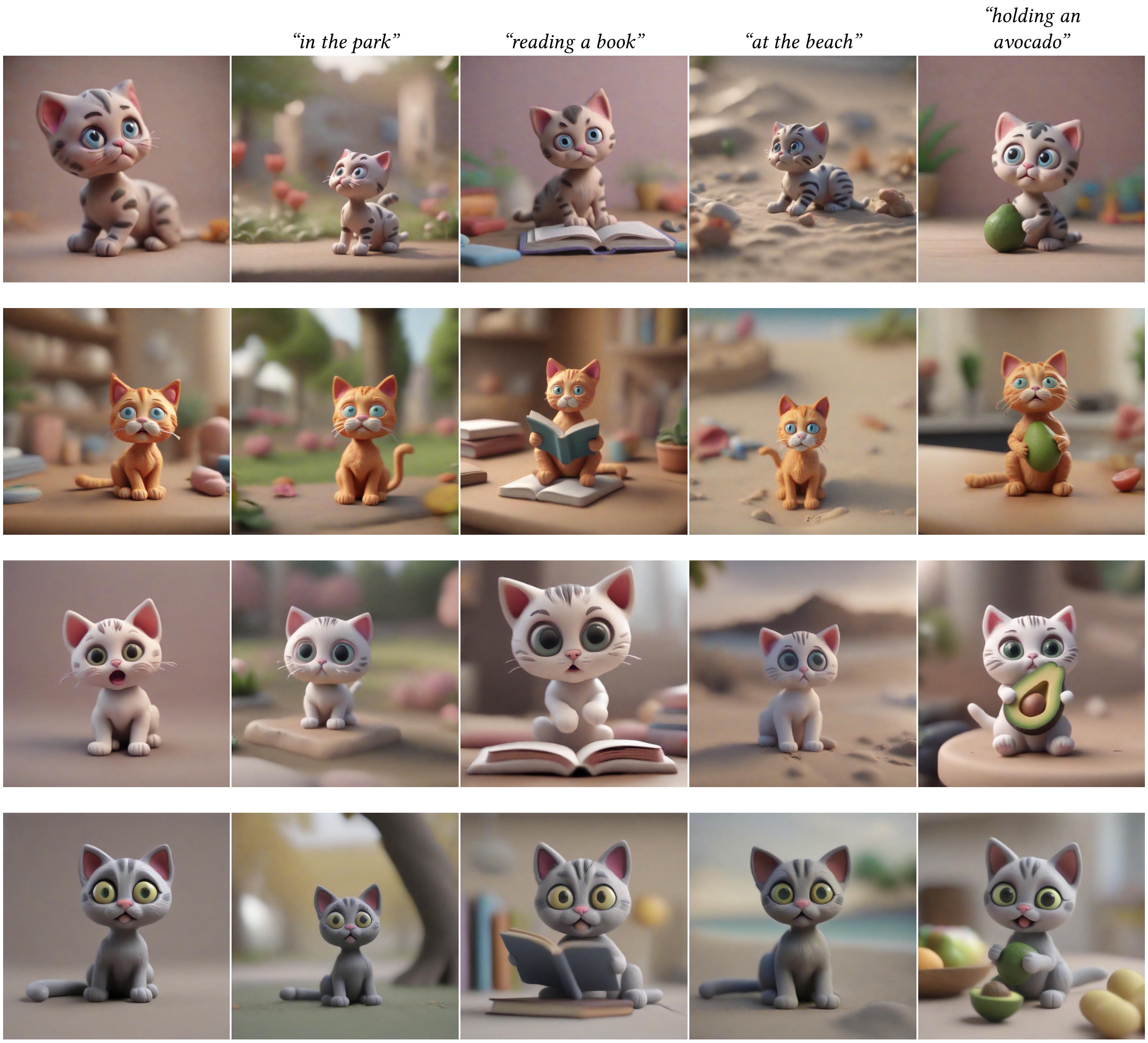


Figure 17: Non-determinism. By running our method multiple times, given the same prompt *"a Plasticine of a cute baby cat with big eyes"*, but using different initial seeds, we obtain different consistent characters corresponding to the text prompt.

Figure 18: Qualitative comparison to naïve baselines. We tested two additional naïve baselines against our method: TI [Gal et al. 2022] and LoRA DB [Ryu 2022] that were trained on a small dataset of 5 images generated from the same prompt. The baselines are referred to as *TI multi* (left column) and *LoRA DB multi* (middle column). As can be seen, both of these baselines fail to extract a consistent identity.

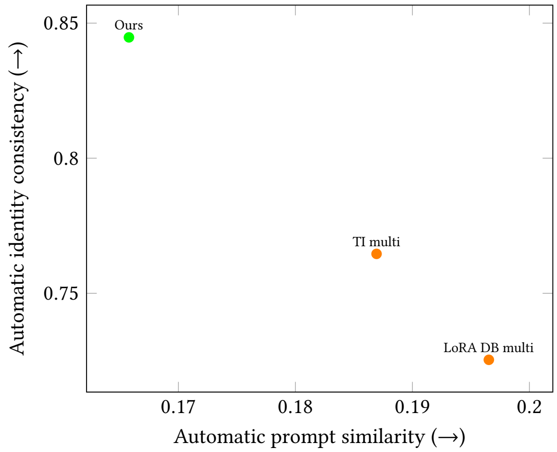


Figure 19: Comparison to naïve baselines. We tested two additional naïve baselines against our method: TI [Gal et al. 2022] and LoRA DB [Ryu 2022] that were trained on a small dataset of 5 images generated from the same prompt. The baselines are referred to as *TI multi* and *LoRA DB multi*. Our automatic testing procedure, described in Section 4.1, measures identity consistency and prompt similarity. As can be seen, both of these baselines fail to achieve high identity consistency.

Figure 20: DALL·E 3 comparison. We attempted to create a consistent character using the commercial ChatGPT Plus system, for the given prompt *"a Plasticine of a cute baby cat with big eyes"*. As can be seen, the DALL·E 3 [Betker et al. 2023] generated characters share only some of the characteristics (e.g., big eyes) but not all of them (e.g., colors, textures and shapes).

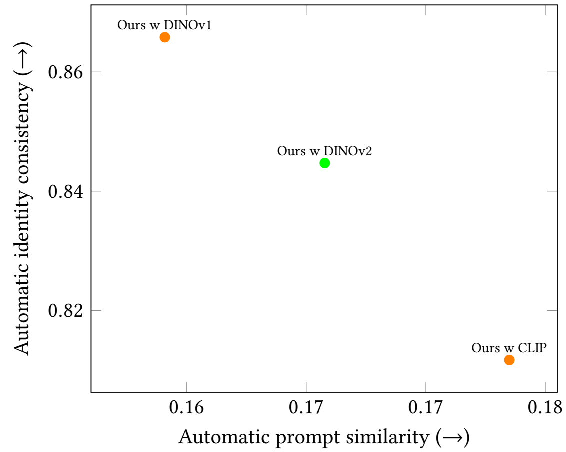


Figure 21: Comparison of feature extractors. We tested two additional feature extractors in our method: DINOv1 [Caron et al. 2021] and CLIP [Radford et al. 2021]. Our automatic testing procedure, described in Section 4.1, measures identity consistency and prompt similarity. As can be seen, DINOv1 produces higher identity consistency by sacrificing prompt similarity, while CLIP results in higher prompt similarity at the expense of lower identity consistency. In practice, however, the DINOv1 results are similar to those obtained with DINOv2 features in terms of prompt adherence (see Figure 22).

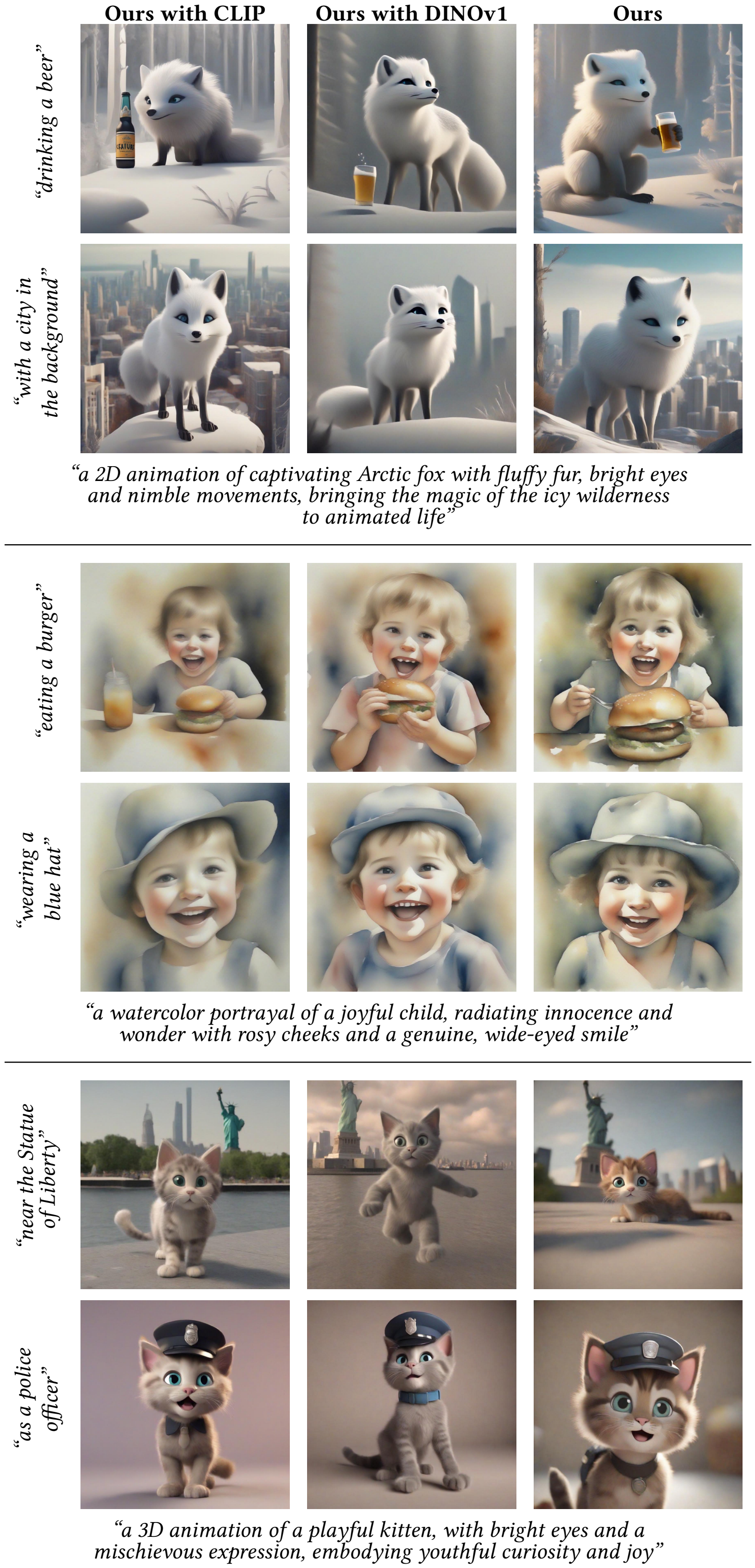


Figure 22: Comparison of feature extractors. We experimented with two additional feature extractors in our method: DINOv1 [Caron et al. 2021] and CLIP [Radford et al. 2021]. As can be seen, DINOv1 results are qualitatively similar to DINOv2, whereas CLIP produces results with a slightly lower identity consistency.

Figure 23: Clustering visualization. We visualize the clustering of images generated with the prompt *"a purple astronaut, digital art, smooth, sharp focus, vector art"*. In the initial iteration (top three rows), our algorithm divides the generated images into three clusters: (1) emphasizing the astronaut's head, (2) an astronaut without a face, and (3) a full-body astronaut. Cluster 1 (top row) is the most cohesive cluster, and it is chosen for the identity extraction phase. In the subsequent iteration (bottom three rows), all images adopt the same extracted identity, and the clusters mainly differ from each other in the pose of the character.

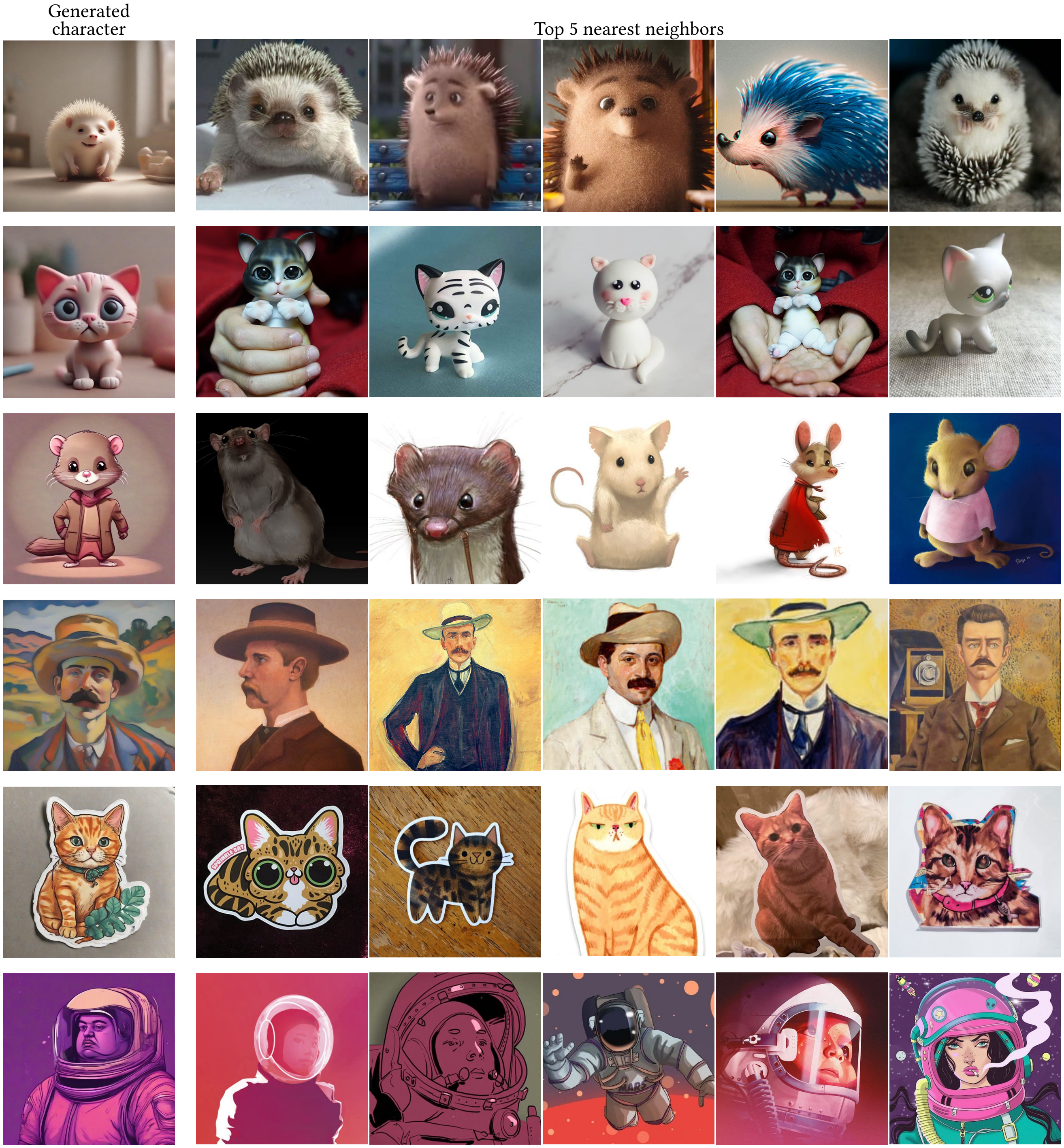


Figure 24: Dataset non-memorization. We found the top 5 nearest neighbors in the LAION-5B dataset [Schuhmann et al. 2022], in terms of CLIP [Radford et al. 2021] image similarity, for a few representative characters from our paper, using an open-source solution [Romain Beaumont 2023]. As can be seen, our method does not simply memorize images from the LAION-5B dataset.

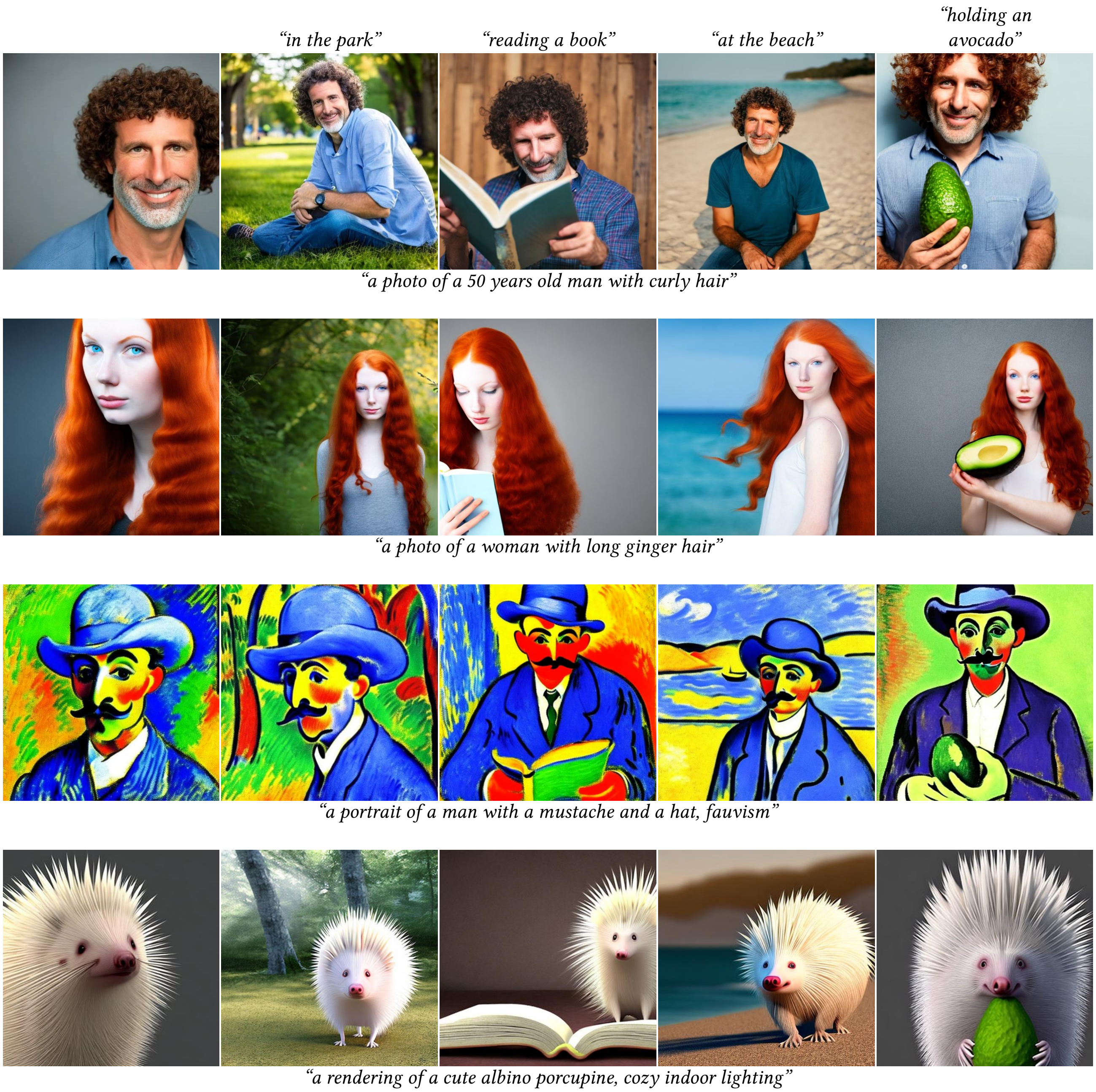


Figure 25: Our method using Stable Diffusion v2.1 backbone. We experimented with a version of our method that uses the Stable Diffusion v2.1 [Rombach et al. 2021] model. As can be seen, our method can extract a consistent character, however, as expected, the results are of a lower quality than when using the SDXL [Podell et al. 2023] backbone that we use in the rest of this paper.