Concept Assets: Naming Custom Concepts in Diffusion Network Outputs

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 1055

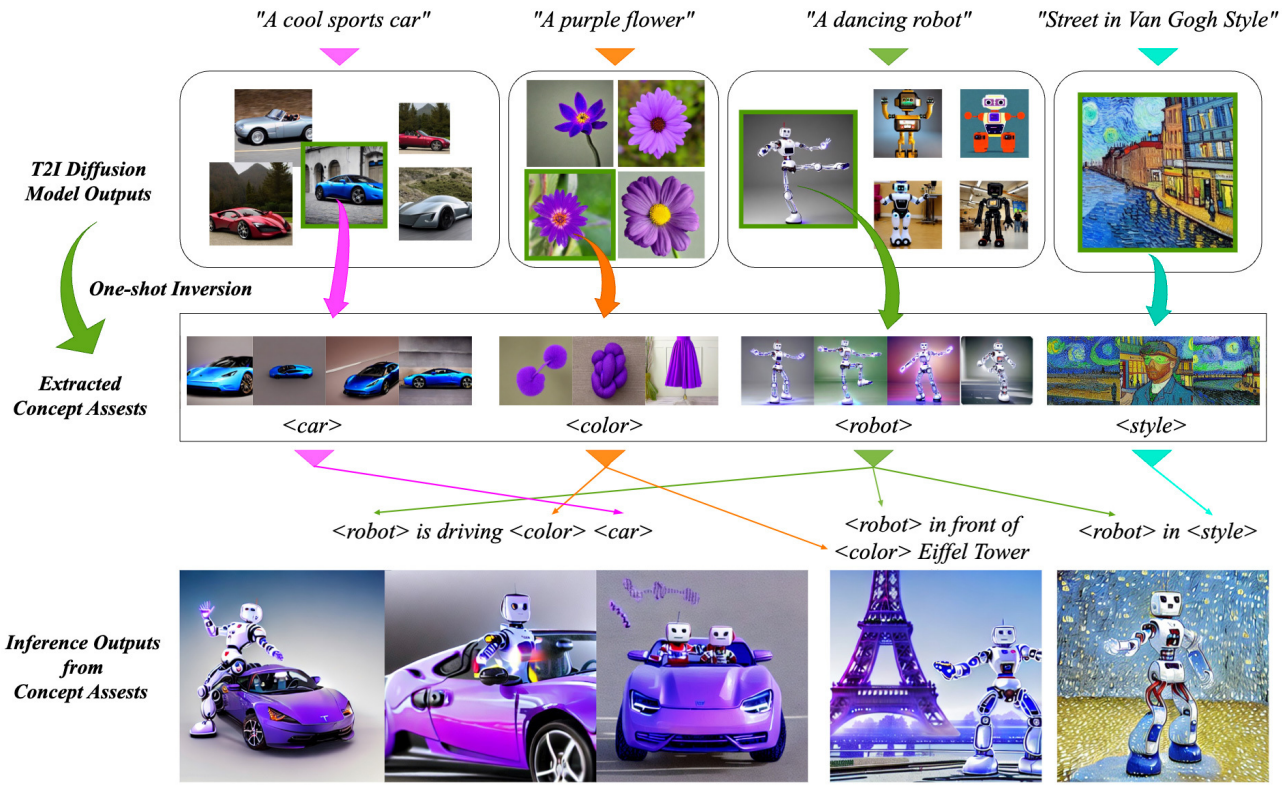

Fig. 1. Concept Assets: The diffusion model takes user-provided text prompts and produces a range of outputs. The user selects their favorite from these (top). Subsequently, we apply our one-shot inversion technique to extract the selected concept(Middle). Finally, these extracted concepts are combined to craft the story of a 'traveling robot' (Bottom).

Recent breakthroughs in text-to-image generation have greatly expanded the possibilities for creating visual content. Despite this progress, these models often struggle to consistently extract not only visual characters but also general concepts, such as style, pose, and color, from pre-trained models. Common approaches typically require multiple pre-existing images of the target subject or fine-tuning with large datasets of target class for one-shot inversion. Our research presents an innovative interactive method for text-driven diffusion models that allows for the extraction and preservation of custom concepts, which we refer to as "concept assets." This mechanism is an augmentation of the textual inversion process, incorporating interactivity that empowers users to edit, define and preserve distinct concepts in generated images for future use, all within a one-shot framework. With this approach, a concept can be captured within a pseudo-token, making it readily available for future use in various applications like character design, story telling and personalized image generation. We quantitatively evaluate our method against existing baselines using multiple metrics and validate our results with a user study. We also provide examples of how our method can be applied in practical scenarios.

CCS Concepts: • **Computing methodologies** → **Image processing**.

Additional Key Words and Phrases: Diffusion customization, concept extraction, generative AI

ACM Reference Format:
Anonymous Author(s). 2024. Concept Assets: Naming Custom Concepts in Diffusion Network Outputs. *ACM Trans. Graph.* 1, 1 (February 2024), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM 0730-0301/2024/2-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

State of art text to image models exhibit remarkable ability to generate convincing images from natural language descriptions; still

there remains an inherent gap between the two domains, which complicates practical usage. Because of the structure and finite quantity of available training data, there is always a degree of ambiguity between a verbal description and appearance, making it difficult to request a particular output in most cases. For example, the phrase *sports car* may correspond to many different images in which the referenced objects are quite diverse in appearance. When generating images for a particular purpose however, the user may wish to retain a particular appearance between iterations to change the context in which a particular object appears, or fine-tune its properties by prompt engineering. But with the exception of very particular terms strongly associated with an appearance in the training data, there is no guarantee that another invocation of the network will yield the same appearance.

A variety of personalization techniques allows the user to associate a special token with the particular appearance of a concept to ensure consistency, but to achieve this they require either multiple pre-existing examples or involved image manipulation such as selection. This limits their utility to concepts of which there are either sufficient pre-existing real world examples, or which are sufficiently concrete to be easily selected.

In this paper, we introduce a new method for personalization which iteratively incorporates the user's feedback to specify a concept that may then be named, saved, and reused in further generation. We extract these concepts from a root image which is either itself generated or provided by the user. This image is automatically modified and the user is then prompted to accept or reject these modifications in order to identify attributes of interest. These attributes can be arbitrary, and include identity (for instance, 'robot'), appearance (e.g. 'purple'), and action (such as 'dancing').

A key aspect of our approach is its ability to disentangle multiple attributes within an image through a combination of pseudoword parameters driven by textual embeddings and LoRA (Low-Rank Adaptation) parameter space. Thus disentangled, attributes from a single root image may then be used and edited independently in future synthesis. For example, we can create variations of a 'dancing robot' with different appearances but the same dance pose. Each resulting variant is then labeled with a descriptive caption, such as 'a <dance1> <robotn>', to highlight the altered attribute. This method of pairing not only aids in distinguishing between attributes but also acts as a self-regulating mechanism in our training process. Through this disentanglement, our approach allows for the efficient extraction of a broad spectrum of concepts in a single training session, each of which is self-consistent. These named attributes, which we dub *concept assets*, then represent concepts of consistent, specific and custom appearance which bridge the domain gap.

We then show that our method can effectively combine both visual and abstract concepts extracted directly from diffusion outputs. This results in output images that feature a natural semantic blend and novel layouts, demonstrating versatility and generalization capability of the extracted assets. Such properties make our method particularly suited for applications like comic creation, storytelling, and poster design.

A key aspect of our approach is that post-extraction, the concept assets are encapsulated entirely within the word embedding of the pseudotokens, and so our technique is not dependent on any particular model architecture or pre-trained weights, and may be transparently combined with any number of techniques which e.g. modify attention weights or the inference process.

In summary, our primary contributions are as follows: (1) We establish a framework for the extraction of multiple general concepts, moving beyond merely generating visually consistent objects. (2) We introduce a lightweight, innovative solution that is non-intrusive to existing architectures and tailored to this expanded scope. It uniquely incorporates user feedback for customization. (3) We conduct comprehensive evaluations, both quantitative and qualitative, including a user study, to demonstrate the robustness and effectiveness of our method in handling a broader range of concepts.

2 RELATED WORK

Text-to-Image Generation Recently diffusion models have driven input-conditioned image generation has undergone a revolution in quality and flexibility. Previously, GAN based methods [6, 17, 32] set a foundation for the field, offering impressive diversity and quality in image generation, yet they often fell short in accurately translating complex textual descriptions into images. Diffusion-based models [8, 11, 13, 15, 26, 28–31] have since made significant progress overcoming this challenge. An increasing number of large-scale text-to-image models [7, 20, 21, 25] are being trained using billions of data sources. These foundation models [4] excel at synthesizing photo-realistic images that correspond to user prompts.

Personalization One of the most impressive aspects of foundation text-to-image models is their capacity for personalization. In this context personalization focuses on consistently generating a user-defined visual concept, bound to a text token, based on a handful of reference images. Two foundational approaches in this area are Textual Inversion (TI) [9] and DreamBooth (DB) [23].

Textual Inversion introduces a unique, trainable text token. This token is optimized to reconstruct the target concept using standard diffusion loss techniques, while the overall model weights remain static. DreamBooth, in contrast, utilizes an existing, but rare, text token. It then fine-tunes the entire model, adapting its weights to better reconstruct the target concept.

Custom Diffusion [16] represents a slightly different approach, fine-tuning only specific layers of the model, whereas LoRA [14, 24] confines its updates to low-rank adjustments. In our research, we employ a two-phase strategy. First, optimize only the textual embeddings of the target concepts. Second, we jointly train both the embeddings and the model weights simultaneously.

Personalization typically involves using target images that contain the same object, to ensure a diversity of outputs. However, extracting concepts directly from diffusion model outputs is more challenging, especially when the desired concept appears only sporadically. Recent developments in this area include one-shot textual inversion techniques like E4T [10], which leverages extensive pre-training across a broad spectrum of concepts within a given domain for quick inversion. Break-a-Scene [1] introduces an innovative approach, utilizing additional mask information and an attention control mechanism for the target visual concept, allowing for more

detailed control over the generated scenes. Like our approach Break-a-Scene [1] uses a similar two-stage training approach.

Despite these advances, methods like E4T and Break-a-Scene face limitations in generalizing to abstract and broad concepts due to their heavy reliance on visual cues.

Another research direction focuses on using text prompts alone to generate consistent concepts, as seen in studies like Elodin [18] and The Chosen One [2]. While these methods achieve visual consistency, their training methodology—expectation and clustering aligned with text prompts—often results in outputs that lack detail and appear overly smoothed. In addition, these methods can only learn a single visually consistent character per text-prompt. In contrast our work enables extracting a consistent character from a single image, giving users much more flexibility.

3 METHOD

The objective of our method is to extract concepts from the diffusion prior by starting from a single seed image, and then augmenting this to a handful of additional images that facilitate learning disentangled concepts. This supports objects and characters, but also the identification of more abstract elements such as actions, colors, styles, and camera views.

In contrast to previous work that relies on training sets of multiple images of the same object or style combined with a fixed user-provided text template, we start from a single root image I and no fixed text template. This image can either be a natural image provided by the user, or one generated by the diffusion model.

Our key insight is to view an image as a mix of different attributes, such as 'identity' (like 'robot') and 'action' (such as 'dancing'). By applying the SDEdit technique with specific text prompts, we are able to alter one attribute while keeping the others constant, creating a variety of images related to the original.

We pair each modified image with a distinct caption that highlights the altered attribute. For instance, if we modify the 'robot' aspect while maintaining the 'dancing' action, the resulting image, denoted as I_n, is then labeled with a caption such as 'a <dance1> <robot N>'. This augmentation process acts as a regularization mechanism which helps correctly disentangled visual attributes. Consequently, a single training yields a variety of pseudo words that can be mixed and matched at inference.

3.1 Pretrained Text-guided Latent Diffusion Models:

Consider a text-guided diffusion model that initiates with a textual embedding C and a random Gaussian noisy image, denoted as I_T. Note that $C = \mathcal{T}(P)$ is the embedding of text prompt P in natural language projected by the text encoder $\mathcal{T}$ and $I_T \in \mathbb{R}^{H\times W\times C}$ is characterized by Gaussian i.i.d pixel values.

The goal of the diffusion model is to progressively denoises the image, resulting in a sequence $I_T, I_{T-1}, \ldots, I_0$ such that I_0 corresponds to the text prompt P. But to allow the diffusion model to operate on a more compact representation, which reduces both time complexity and memory usage, Latent Diffusion Models (LDMs) [22] uses an encoder $\mathcal{E}$ to map a given image I into a latent embedding z. Subsequently, a decoder $\mathcal{D}$, is employed to reconstruct the input image from z, such that $\mathcal{D}(\mathcal{E}(I)) \approx I$. Therefore, we only need to replace the image I with its latent embedding z in the following algorithm.

z_t is a sample with added standard Gaussian noise ϵ, where the noise is introduced according to a time-dependent schedule α_t. The relation is given by $z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon$, corresponding to a forward diffusion process:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I) \tag{1}$$

To learn the reverse process, the mean value $\mu_\theta(x_t, t)$, parameterized by θ, is predicted:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta) \tag{2}$$

A denoiser module, parameterized by the network ϵ^θ, is utilized for fitting the objective, which is equivalent to fitting the mean value prediction in the reverse process:

$$\min_\theta \mathbb{E}_{z_0,\varepsilon\sim\mathcal{N}(0,I),t\sim\text{Uniform}(1,T)} \|\varepsilon - \varepsilon_\theta(z_t, t, C)\|^2. \tag{3}$$

During the inference phase, the model can employ either stochastic sampling as in DDPM [12], which introduces noise at each sampling step, or deterministic sampling as in DDIM [27], which follows an ODE-like deterministic trajectory. Specifically, we utilize the deterministic DDIM sampling approach to leverage the inverse characteristics of the ODE path:

$$\frac{z_{t-1}}{\sqrt{\alpha_{t-1}}} = \frac{z_t}{\sqrt{\alpha_t}} + \left(\frac{\sqrt{1-\alpha_{t-1}}}{\sqrt{\alpha_{t-1}}} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\right)\epsilon^\theta(z_t, t, C) \tag{4}$$

3.2 Text Guided SDEdit for Data Augmentation:

In our approach, we adapt the Stochastic Differential Equation (SDE) editing technique, which introduces noise into an existing image and then partially reverses this process to produce a new image. Originally, this technique was used for editing real-world images, using manual strokes for edits, and the subsequent denoising step did not utilize textual prompts. However, we have modified the method to better suit our objectives. Our adapted procedure involves introducing noise into the image and then employing text-guided, classifier-free guidance $\epsilon^\theta(z_t, t, C)$ during the DDIM denoising phase to achieve the desired outcome.

Initially, the input image is converted into a latent space representation through an encoder, denoted as $z = \mathcal{E}(I)$. This latent representation, z, is then subjected to the addition of independent standard Gaussian noise, expressed as $z_t = \sqrt{\alpha_t}z + \sqrt{1-\alpha_t}\epsilon$. The denoising process commences not from the initial step T but from an intermediate step t. Note that if noise is added too early in the sequence, specifically at a timestep near T, this can lead to newly generated images with no relation to the root image I. if the noise is introduced at a timestep close to 0, which is too late in the sequence. Such a timestep can result in insufficient variations in the content of the newly generated images. Therefore we choose to introduce noise at timestep $T = 0.86$ to balance variety and identity preservation.

3.3 Interactive User Selection for Dual images and Captions:

Based on their goals users can use text-guided SDEdit to generate a wide range of variations from the initial seed image; then select their preferred variations to use when learning disentangled concepts.

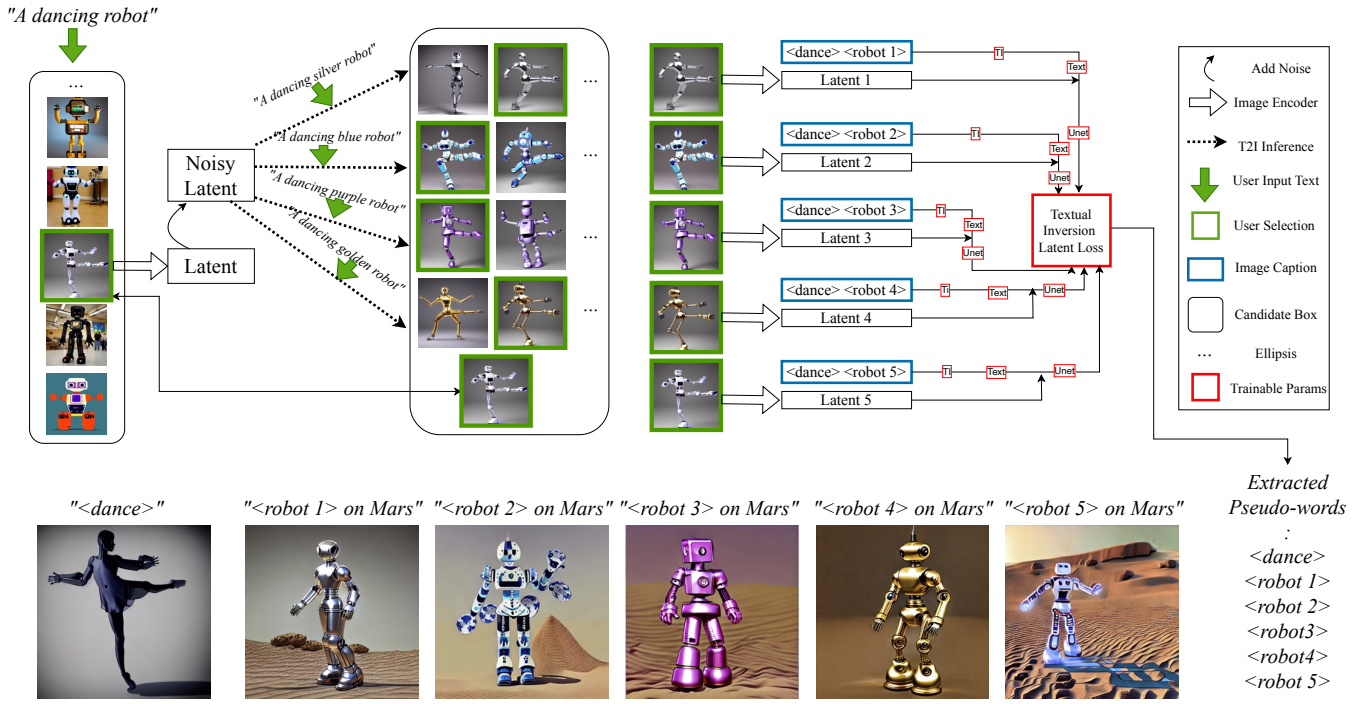

Fig. 2. Overview of the Method: Our approach is divided into three primary phases. (1) From the user's text prompt, an initial image is chosen as the starting point of the process from among various outputs, ensuring it closely aligns with the user's requirements. (2) We then employ text-guided image perturbation methods to create a range of potential images. The user selects a group of images that exhibit a common characteristic, such as a similar 'dance' pose. (3) Each chosen image is then labeled with a simple caption that highlights its distinct, intrinsic attribute. Following this, we use textual inversion and LoRA training to simultaneously extract multiple pseudo-concepts (or concepts).

This selection process can be iteratively refined by user feedback, maintaining image quality without introducing degradation or artifacts. This interactive step is crucial for users to tailor the semantic details they desire from the pre-trained model, as shown in Figure 2.

For instance, if the original image I is categorized as a 'dancing robot', it represents a combination of the action 'dance' and the identity 'robot'. To isolate the 'dance' action and identify the 'robot', we can use SDEdit with prompts like 'a dancing silver robot', 'a dancing purple robot', 'a dancing single-eyed robot', etc. Each prompt generates a batch of images from which users can choose. The chosen image can either be added to the training dataset or returned to the SDEdit process for further refinement. For our training set, we select eight images based on empirical findings. The effects of using fewer or more images in the training set are shown in Section 4.3. Each training image is labeled with a unique caption, such as '<dance 1> <robot 1>', through to '<dance 1> <robot N>'. While we vary the robot identities, we keep the dance action constant, resulting in a single action pseudoword '<dance 1>' to represent the common dancing pose, and multiple identity pseudo-words to distinguish the different robot appearances.

Figure 3 showcases various use cases, demonstrating how our method allows for the creation of desired image variations based on textual guidance provided by users. Utilizing our approach, these varied images can be converted into pseudo-words which can be invoked for subsequent text-to-image inference tasks. Thus, the variation generation stage is a valuable tool for users to directly design and refine concepts. In Figure 8, we display the results of inference using these extracted and combined assets.

3.4 Multiple Words Textual Inversion:

Our methodology is anchored in the application of the pre-trained Stable Diffusion [21] model, adapted for our specific purposes. Consistent with the textual inversion framework, we extend the tokenizer's lexicon by incorporating each pseudo word $< S_i >$, represented as a novel word embedding v_i within the text encoder. Our approach diverges from prior formulations by learning a set of parameters $V = \{v_i\}_{i=1}^{n}$ concurrently. Moreover, we enhance the consistency and expressive capacity of the pseudo words by updating the residual weights of the model ΔW. This update is effected through a low-rank adaptation (LoRA) applied to both the text-encoder and the U-Net component of the architecture.

The optimization objective is formulated thus:

$$\Theta = \arg\min_{\Theta=\{V,\Delta W\}} \mathbb{E}_{z\sim\mathcal{E}(x),y\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\Theta(z_t^k, t, c_\Theta^k(y))\|_2^2\right], \tag{5}$$

$$k = 1, 2, \ldots, n \tag{6}$$

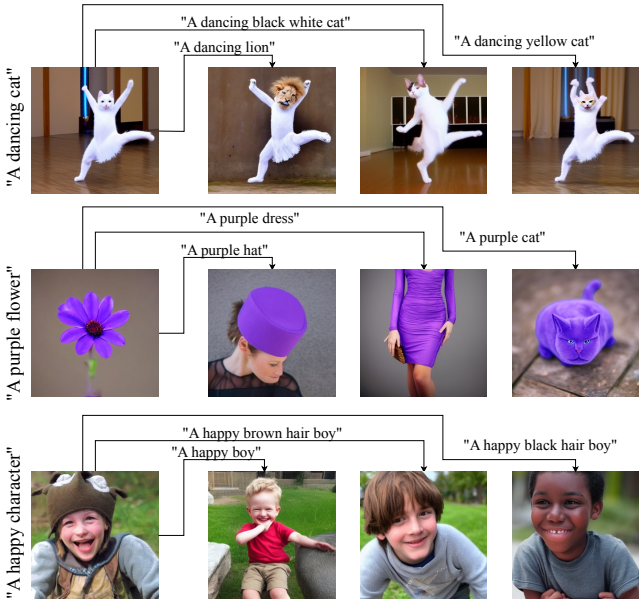

Fig. 3. User-Driven Conceptual Editing Phase: This stage begins with the images selected by the user. Users provide text prompts for various alterations. As this editing process relies solely on text, it enables the creation of more diverse concepts from the original image with ease. The provided examples demonstrate that the edits can range from subtle (altering only the head style in the 'cat' example) to imaginative (creating a 'Ballet Lion'). The results effectively retain key characteristics from the original image, such as the consistent purple color seen in the second row and consistent happy mood seen in the third row. Additionally, the edits blend seamlessly into the natural image distribution, as seen in the perturbed 'boy face' images, which appear natural and undistorted.

In this expression, y^k corresponds to the caption of the k-th image, and z_t^k is the perturbed latent variable of the k-th image, injected with noise according to the timestep t. It is important to emphasize that the pairing of images and captions is fixed, as delineated in the previous section. Each selected image I_k is paired exclusively with a single caption $' < A_1 >< B_k >'$, where A and B represent two distinct characteristics that inherently differentiate the various elements within the image's composition.

Note that in the second stage of our approach, we fine-tune not only the textual embedding but also text-encoder and the U-Net component of the stable-diffusion model. During inference, especially when incorporating multiple custom pseudo words into a single text prompt, it's necessary to integrate more than just the textual embeddings of these pseudo words. Specifically, we also need to combine their LoRA matrices. When adding the delta parameterized by a LoRA matrix, apply a scaling factor of 0.8. In cases where multiple LoRA meetings (from multiple concepts) are invoked in a single training, we scale the first by 0.8 and subsequent ones by 0.5 (order determined by pseudo-word order in text prompt). For a more comprehensive understanding of the LoRA fine-tuning method, readers are encouraged to consult the referenced citation[14][24]."

4 EXPERIMENTS

In Section 4.1, we evaluate our method against various baselines through qualitative and quantitative comparisons. Section 4.2 describes the user study along with its findings. In Section 4.3, we present the results of our ablation study. Lastly, Section 4.4 showcases several practical applications of our method.

4.1 Quantitative and Qualitative Comparison

For a quantitative evaluation of our method, we compute the text-image and image-image feature similarity to numerically measure the prompt consistency and identity consistency, following [3]. First, we automatically generate 10 candidates using ChatGPT for the image character and the text description, similar to [3]. Then, we compute the text-image feature similarity as a prompt consistency metric by using the CLIP [19] text and image encoder to calculate the feature distance between the corresponding prompt and the generated image. For the image-image similarity, CLIP image feature embedding similarities are measured for the generated images across the same prompt concept as the identity consistency.

The left plot in Figure 5 visualizes the result of the quantitative evaluation of baselines and our models in a 2D plot where the x-axis is prompt consistency and y-axis is identity consistency. Our model achieves the best prompt consistency with relatively good identity consistency compared to the baselines, LoRA DB [24], Textual Inversion (TI) [9], The Chosen One [2] and Break-A-Scene [1]. We have observed that baselines with a relatively high identity consistency score, such as LoRA DB, actually produce candidates with limited diversity. In contrast, our method maintains a sweet spot between identity consistency and generative diversity.

A visual comparison with the baselines is shown in Figure 4. Our model is able to generate results that are consistent to the text prompt while maintaining the original object identity (<robot> and <car>). Furthermore, our model can accurately extract abstract concepts like <style> in the last two rows of Figure 4.

4.2 User Study

To thoroughly evaluate our model based on a subjective assessment by human participants, we conducted a user study to gauge the consistency rate of the generated samples. Users were asked to rank each candidate from 1 to 5 for all five methods by independently considering the prompt consistency and identity consistency. We used the same platform, Amazon Mechanical Turk (AMT), to distribute the studies to participants. After averaging the scores across all categories, we obtained the final score, which is presented as the user study score, shown in the right plot in Figure 5.

4.3 Ablation Study

We performed an ablation study to determine the optimal number of variation inputs for training identity concept inversion. As shown in Figure 6, the results, indicate that training with a single image often leads to overfitting, resulting in the generation of images with limited diversity and compromised image quality during inference. Conversely, training with an excessive number of pseudo words (up to 16) can cause the model to overfit, which negatively impacts the diversity and fidelity of the background content. To ensure a

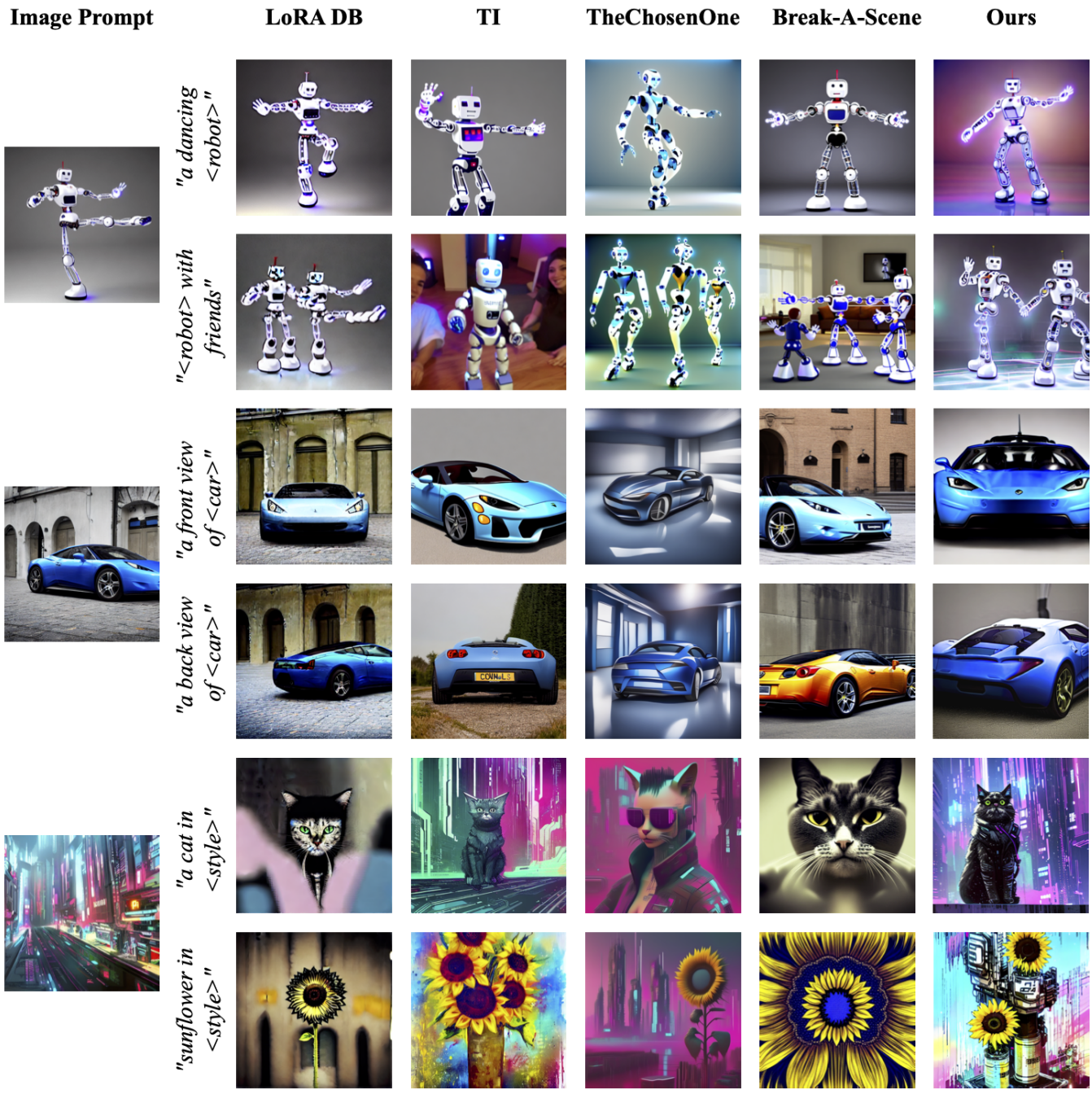

Fig. 4. Qualitative Comparison. In this comparison, we evaluate our method against several baselines: LoRA DreamBooth[24] and Textual Inversion[9], both of which use a single image as input. TheChosenOne[2] does not require an image prompt for learning inversion; instead, we use ChatGPT to generate a detailed image description and input this into TheChosenOne for concept generation. We also provide a detailed mask to Break-A-Scene[1] for concept extraction. The qualitative results indicate that LoRA DreamBooth[24] tends to overfit. Textual Inversion[9] and Break-A-Scene[1] do not accurately learn the concept from the image prompt. Although the concept convergence in TheChosenOne[2] is consistent, it exhibits a style bias. Conversely, our method successfully follows the prompt and maintains consistent identities, while generating characters in varied poses and viewing angles.

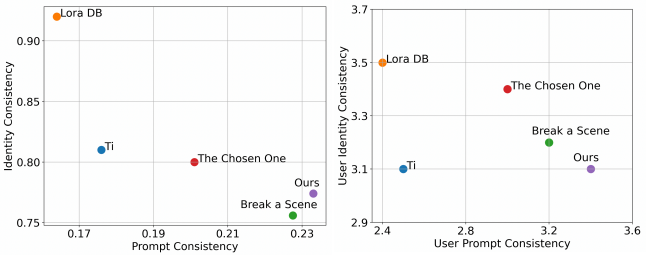

Fig. 5. Quantitative result (left) and User study (right) comparing with various baselines. We show that our method can get a consistent favorable choice under both feature similarity and user choice under random choice prompts and characters.

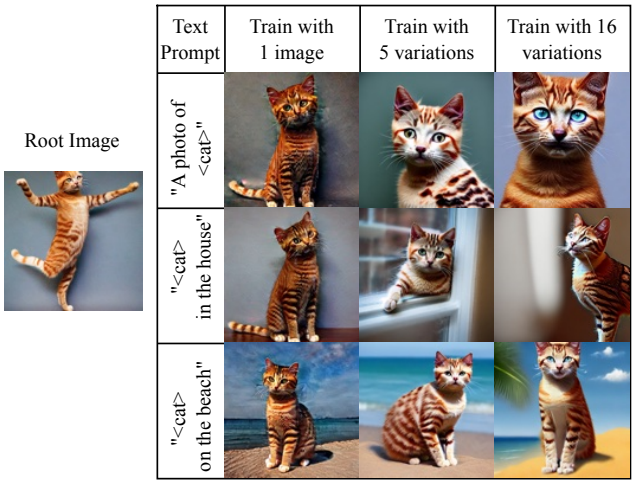

Fig. 6. Ablations on number of training variations

balanced comparison, we varied the number of textual inversion training steps in conjunction with LoRA training steps (500+500, 1000+1000, 1500+1500) and selected the best visual outcomes for the figure. Based on our observations, employing 5-8 variations and 1000+1000 steps within our framework yielded the most effective results in terms of concept consistency, prompt alignment and overall image quality.

4.4 Applications

4.4.1 General Concept Extraction. Figure 8 presents various disentangled concepts derived from the examples in Figure 3. A key observation is that abstract concepts, although not visually explicit like 'Ballet', 'Color', and 'Happy', can be successfully extracted and applied to a variety of identity concepts. For additional examples demonstrating how our extracted concepts can be generalized and combined with various textual guidance during inference please see the Appendix.

4.4.2 Multiple Pseudowords Combination. Figure 1 illustrates the integration of multiple extracted assets from different experiments, showcasing the versatility of our method in applications such as storytelling, comic book creation, poster design, and more. However, we find that combining an excessive number of concepts (more than four) in a single image is challenging. Additionally, we note that concepts related to style and color are more readily combined compared to visual identities. These limitations are further discussed in Section 5.

5 LIMITATIONS AND CONCLUSIONS

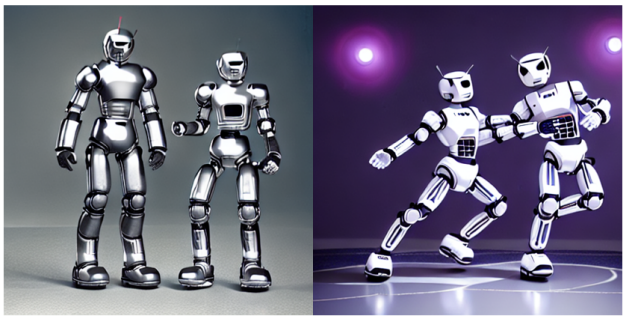

Fig. 7. Limitations examples.

Because our method does not modify the underlying diffusion model in any way, it exhibits the same issues those models do. This includes problems like catastrophic neglect and incorrect attribute binding, described in prior literature [5]. This manifests as e.g. inability to generate two concepts from the same class during inference, particularly when these concepts are textually similar, as illustrated in Figure 7. However, the same prior work describes solutions to this problem and may be combined with our method, again because we do not modify the underlying diffusion model and so retain compatibility with similar enhancements.

In conclusion, our method presents an innovative interactive method for text-driven diffusion models, significantly enhancing the textual inversion process to capture and retain unique 'custom assets.' This approach is user-friendly, facilitating effortless editing and application of these concepts in diverse areas such as character design and personalized image generation, all within a streamlined one-shot framework. Our method surpasses the limitations of merely achieving visual consistency, introducing a lightweight and adaptable solution that effectively incorporates user input without altering the underlying model architecture. Extensive evaluations, including comparisons with existing baselines and a user study, have validated the efficacy and adaptability of our method in managing a broad spectrum of concepts.

REFERENCES

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-A-Scene: Extracting Multiple Concepts from a Single Image. *arXiv preprint arXiv:2305.16311* (2023).

[2] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. The Chosen One: Consistent Characters in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2311.10093* (2023).

[3] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. The Chosen One: Consistent Characters in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2311.10093* (2023).

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[5] 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 Trans. Graph.* 42, 4, Article 148 (jul 2023), 10 pages. https://doi.org/10.1145/3592116

[6] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583* (2022).

[7] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. 2023. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807* (2023).

[8] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.

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

[10] 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.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[13] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. 2023. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093* (2023).

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. [n. d.]. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*.

[16] 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.

[17] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. 2021. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573* (2021).

[18] Rodrigo Mello, Filipe Calegario, and Geber Ramalho. 2023. ELODIN: Naming Concepts in Embedding Spaces. *arXiv preprint arXiv:2303.04001* (2023).

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

[20] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

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

[23] 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.

[24] Simo Ryu. 2023. Low-rank adaptation for fast text-to-image diffusion fine-tuning.

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).

[26] 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.

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[28] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. 2021. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems* 34 (2021), 1415–1428.

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

[30] Yang Song and Stefano Ermon. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems* 33 (2020), 12438–12448.

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. [n. d.]. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

[32] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. 2021. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792* (2021).

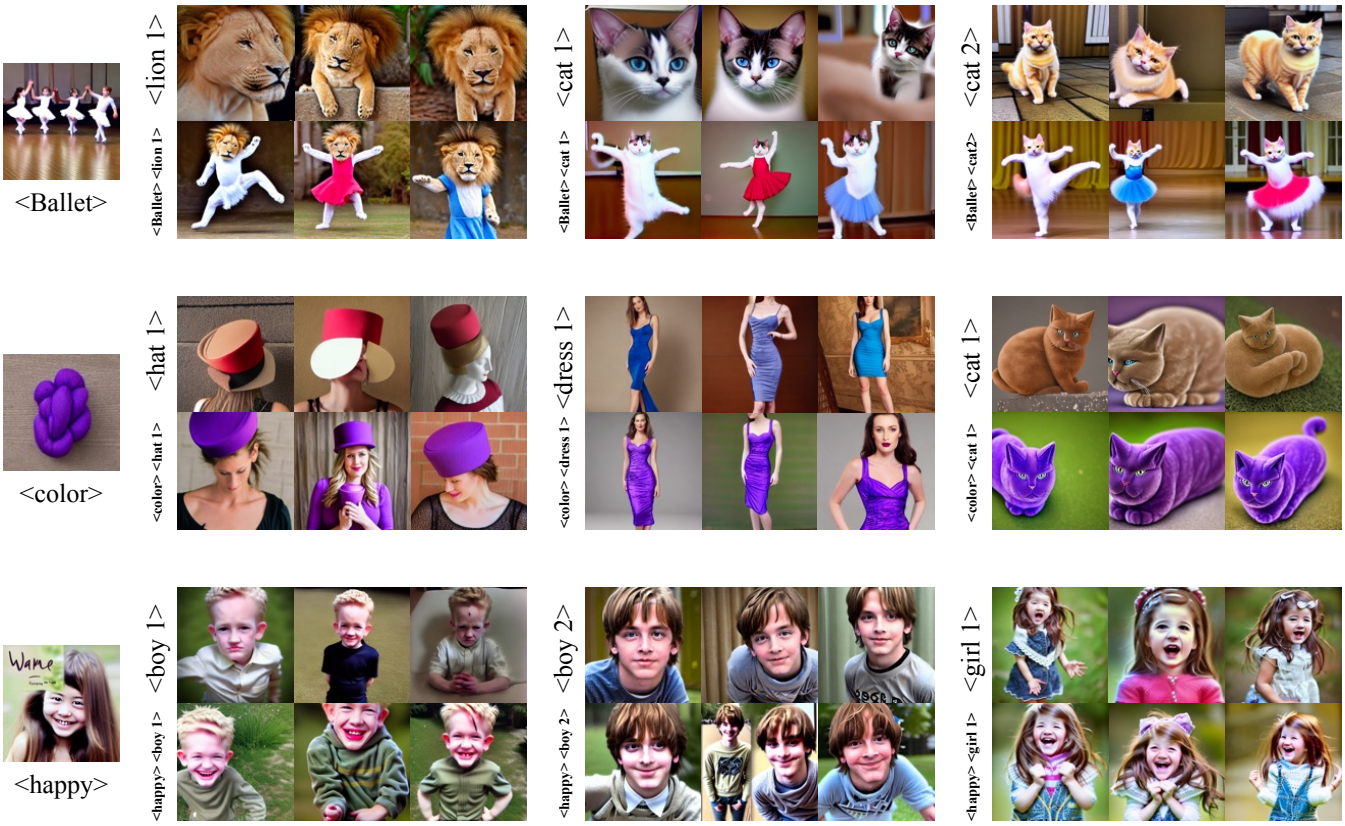

Fig. 8. General Concepts

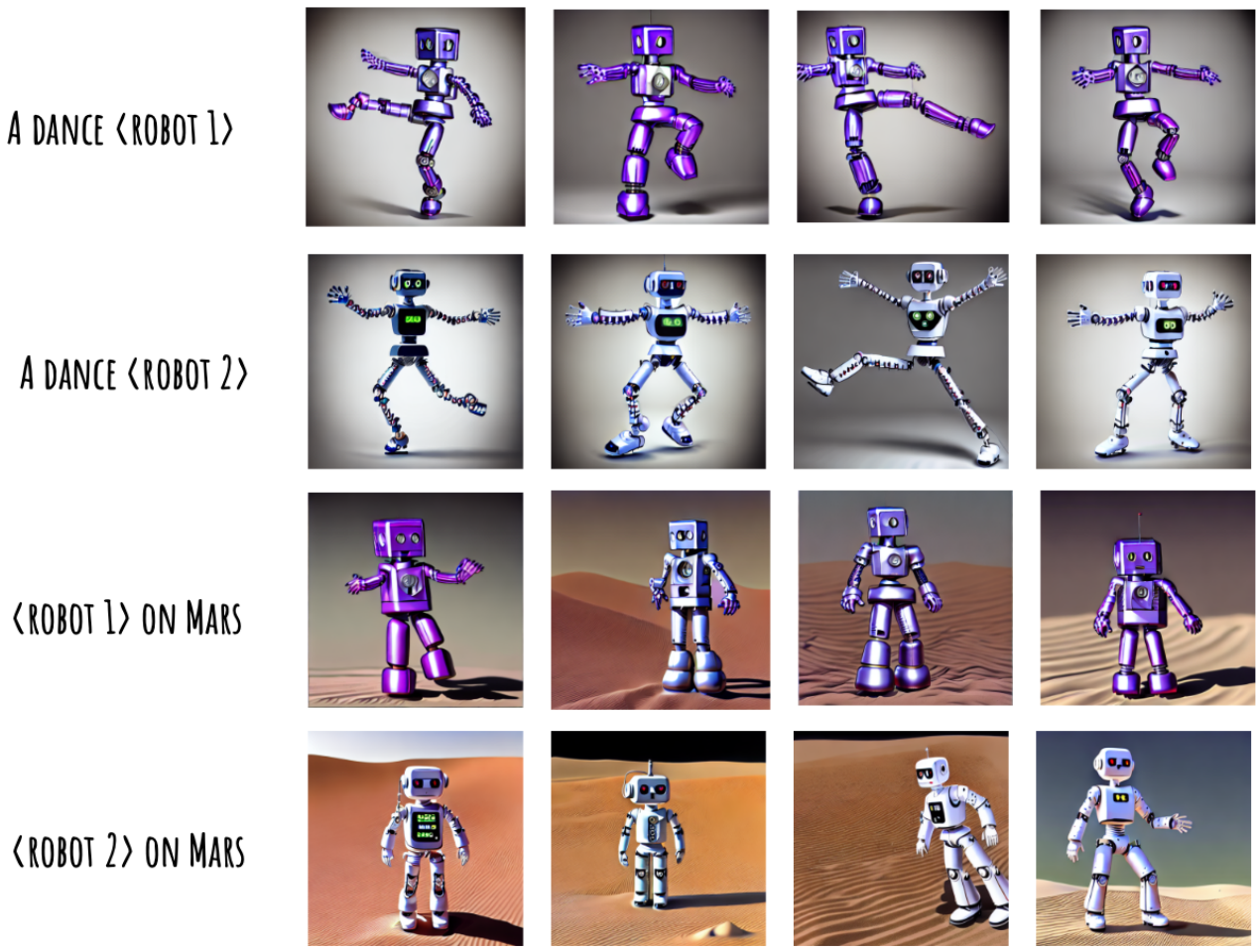

Fig. 9. More visualization examples, we show our method can generalize in different cases.

Fig. 10. More visualization examples, we show our method can generalize in different cases.

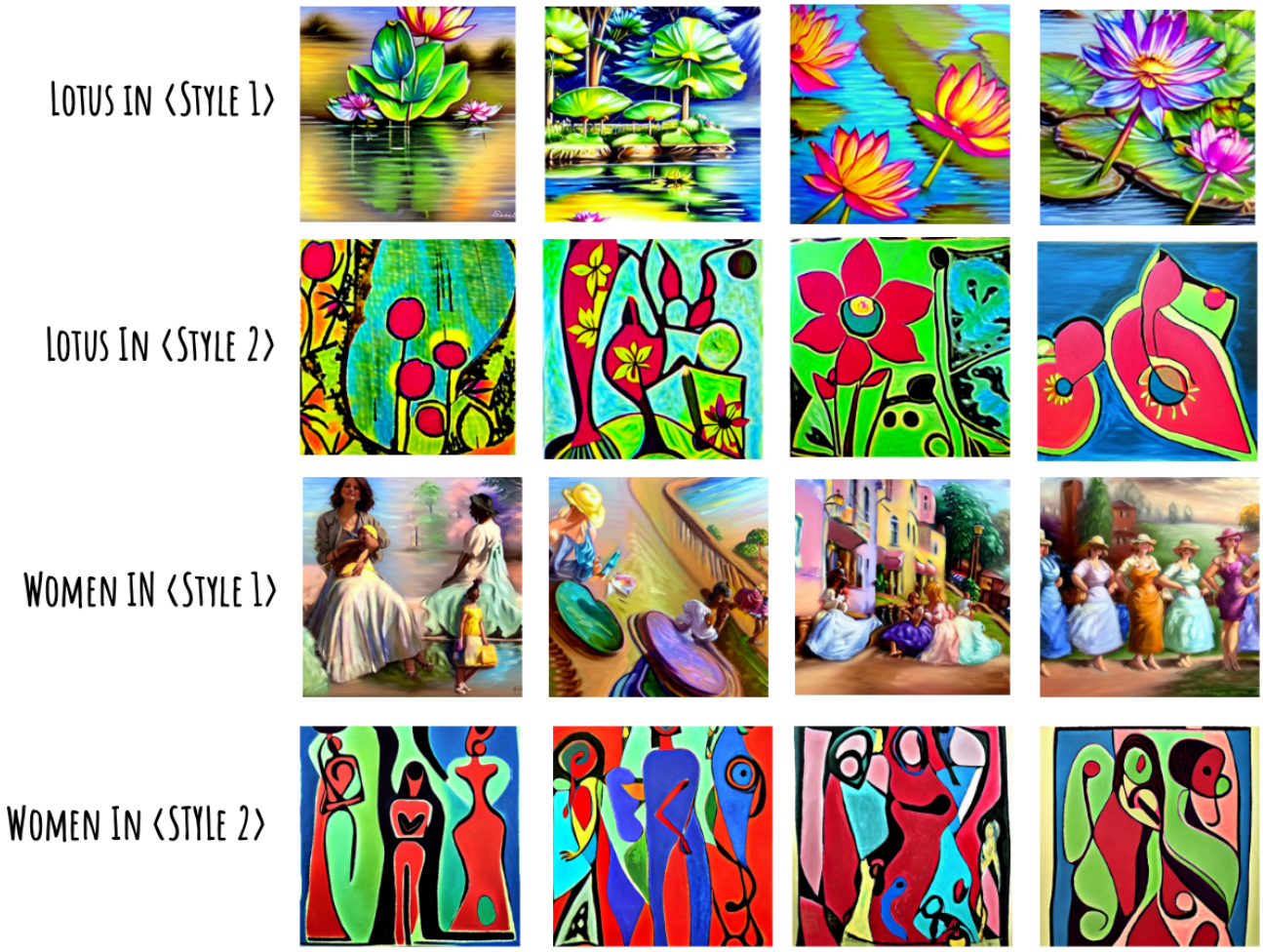

Fig. 11. More visualization examples, we show our method can generalize in different cases.