Synthetic-driven Text-to-Image Generation with Preference-based Reinforcement Learning

Bumkyu Choi, Namhoon Jung and Kyutae Cho

LIGNEX1 CO., Ltd
{bumkyu.choi, namhoon.jung, kyutae.cho2}@ligenex1.com

Abstract

Most existing subject-driven text-to-image generation methods predominantly utilize real-world images as their reference data. However, in fields where obtaining real data is difficult or impossible such as military applications or speculative futuristic concepts, this approach is quite limited. In this work, we address this challenge by using virtually generated reference data, such as computationally synthesized designs or conceptual prototypes, as the starting point for subject-driven image generation. By leveraging these artificial references with preference-based reinforcement learning, we can flexibly produce new, scenario-specific visuals that align with the intended applications and requirements. Our experiments suggest that this approach could enables the creation of highly relevant datasets for tasks such as object detection, and recognition highlighting the potential of virtually seeded data pipelines for practical deployment.

Ⅰ. INTRODUCTION

Recent advances in subject-driven text-to-image generation, such as DreamBooth [1] and Subject-Driven Text-to-Image (SuTI) [2, 3], have enabled users to generate highly customized images from textual prompts and reference images. However, these methods are often grounded in the availability of real-world reference data, which presents a major limitation in domains where data acquisition is constrained or impractical. For instance, sensitive areas such as the military or speculative fields involving futuristic or conceptual designs often lack access to sufficient real images.

To address this gap, we propose a new approach that utilizes virtually generated reference data, produced via generative models and conceptual design, as a starting point for subject-driven image generation. Leveraging synthetic references and fine-tuning with preference-based reinforcement learning, our approach supports the creation of diverse and scenario-specific datasets even in the absence of real-world imagery.

This virtually seeded generation pipeline extends the applicability of personalized generation techniques to previously inaccessible domains. While we do not directly evaluate downstream tasks, the semantic quality and consistency of the generated images suggest their potential utility in applications such as object detection or recognition. Extensive experimentation highlight the practicality and adaptability of the proposed method for dataset creation in challenging or novel environments.

Ⅱ. RELATED WORKS

Ⅱ-1. Subject-Driven Text-to-Image Generation

Subject-driven text-to-image generation aims to produce images that accurately reflect a specific subject based on textual prompts and reference images. Previous works like DreamBooth [1] and SuTI [2, 3] enable personalized image generation by fine-tuning diffusion models with a small number of real images to preserve subject identity in diverse contexts. However, a common limitation among these approaches is their reliance on real-world reference data, which restricts their applicability to domains where acquiring such data is challenging or unfeasible.

Building upon these methods, we focus on extending their applicability to scenarios lacking real-world reference images by leveraging synthetic virtual data and preference-based reinforcement learning. We employ a two-stage pipeline, where the first stage involves generating reference images, and the second stage utilizes these generated images as training references.

Ⅲ. METHODS

Ⅲ-1. Reference Images Generation

To generate virtual reference images, we utilized recent advanced generative models including Stable Diffusion (XL, 3.5), and the GPT-Image Generation model. The objective was not merely to produce high-quality individual images, but rather to ensure

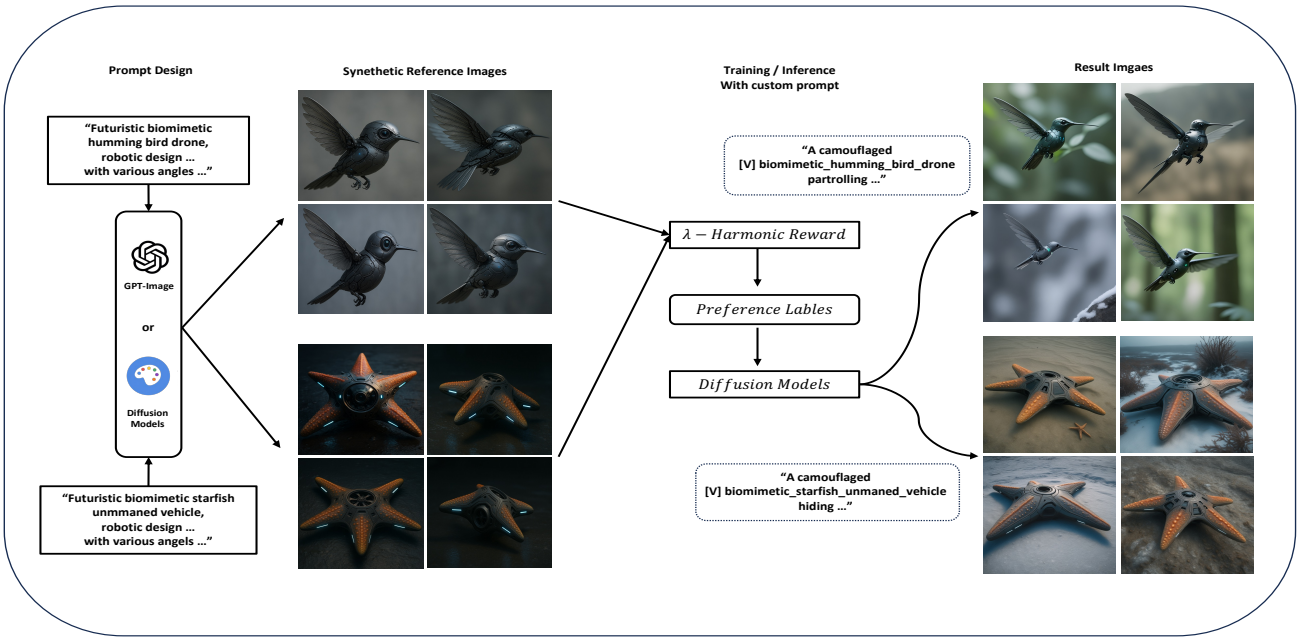

Figure 1. The proposed framework generates subject-driven images using synthetic virtual references and preference-based reinforcement learning. Illustrated examples include biomimetic hummingbird drones and starfish-shaped unmanned vehicle.

consistency across multiple viewpoints, creating cohesive multi-angle visualizations of each conceptual design. As practical examples, we applied this generation strategy to futuristic biomimetic designs such as 'biomimetic hummingbird drones' and 'biomimetic starfish unmanned vehicle'.

Ⅲ-2. Fine-tuning with RPO

In the training phase, we applied Reward Preference Optimization (RPO) [3] to fine-tune the model using a limited set of virtual reference images. RPO leverages a reward function that promotes alignment between image and text representations, allowing effective fine-tuning without relying on human-labeled feedback. This approach remains robust even when synthetic references are used in place of real-world image datasets.

As illustrated in **Figure 1.** we can use these artificially generated images, which do not exist in the real world, to create various scenarios. Moreover, our two-stage pipeline eliminates the need for meticulously crafted prompts; instead, specialized image generation can be guided using simple, unique tokens.

Quantitative evaluation yielded the following similarity scores (synthetic reference / generated result): Biomimetic hummingbird — **DINO**: 0.888 / 0.699, **CLIP-I**: 0.961 / 0.833, Biomimetic starfish — **DINO**: 0.796 / 0.682, **CLIP-I**: 0.922 / 0.867

IV. CONCLUSION

This paper introduces a practical and adaptable framework for text-to-image generation targeting domains where real-world reference images are difficult or impossible to obtain. By synthesizing semantically coherent reference images from text, our proposed pipeline enables subject-driven generation in fields where the target concepts are hypothetical, conceptual, or beyond current real-world observation. Future work may explore the integration of these high-quality synthetic datasets into downstream pipelines, such as object detection, and recognition to further assess their practical utility.

REFRENCES

[1] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 22500-22510 (2023).

[2] W. Chen, H. Hu, Y. Li, N. Ruiz, X. Jia, M.-W. Chang, W. W. Cohen, "Subject-driven text-to-image generation via apprenticeship learning," Advances in Neural Information Processing Systems, vol. 36, pp. 30286-30305 (2023).

[3] Y. Miao, W. Loh, S. Kothawade, P. Poupart, A. Rashwan, Y. Li, "Subject-driven text-to-image generation via preference-based reinforcement learning," Advances in Neural Information Processing Systems, vol. 37, pp. 123563-123591 (2024).