Less-to-More Generalization: Unlocking More Controllability by In-Context Generation

Shaojin Wu Mengqi Huang* Wenxu Wu Yufeng Cheng Fei Ding† Qian He
Intelligent Creation Team, ByteDance
{wushaojin, huangmengqi.98, wuwenxu.01, chengyufeng.cb1, dingfei.212, heqian}@bytedance.com

Project Page: https://bytedance.github.io/UNO

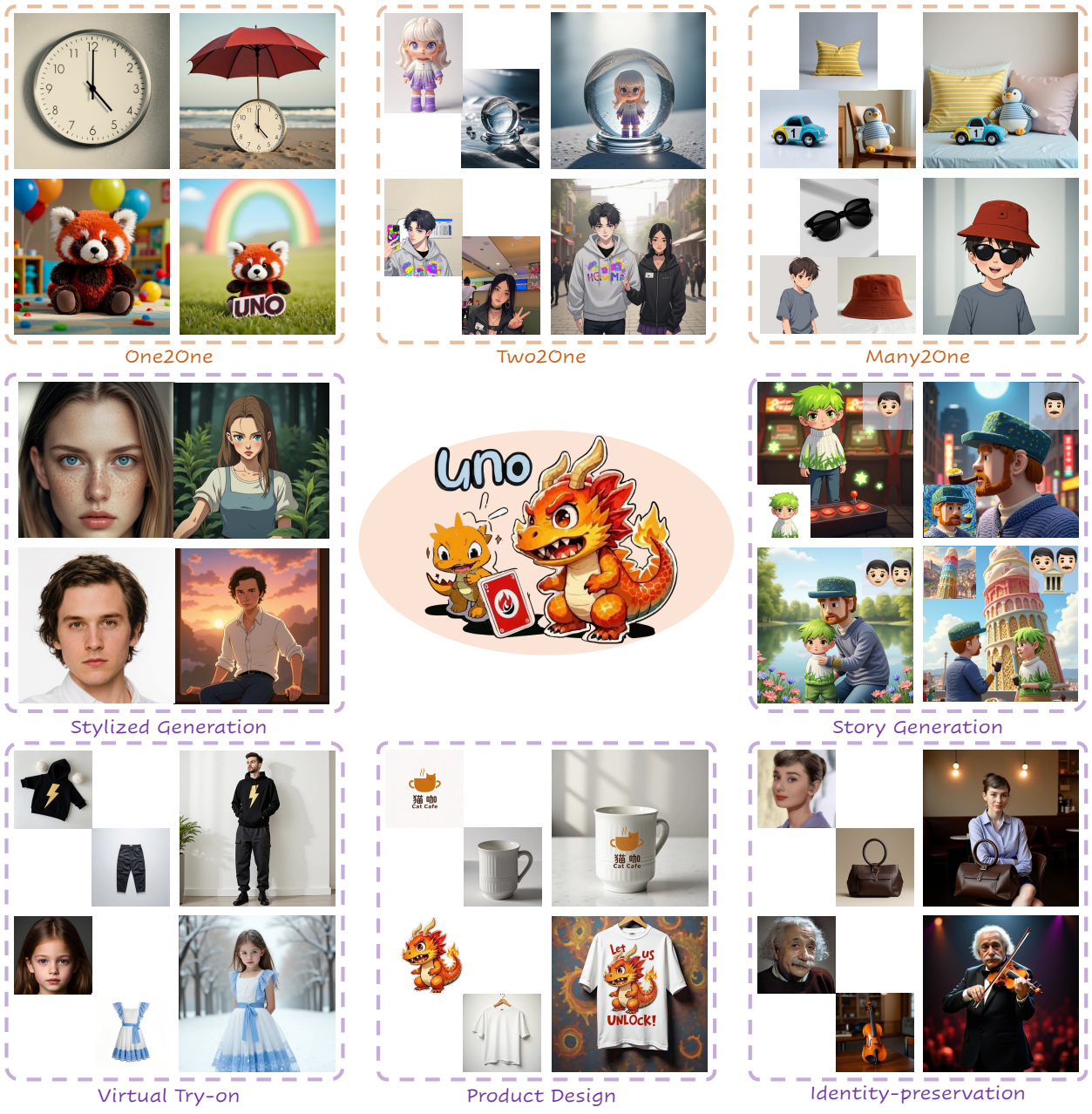

Figure 1. Our UNO evolves as an universal customization from single to multi-subject.

Abstract

*Although subject-driven generation has been extensively explored in image generation due to its wide applications, it still has challenges in data scalability and subject expansibility. For the first challenge, moving from curating single-subject datasets to multiple-subject ones and scaling them is particularly difficult. For the second, most recent methods center on single-subject generation, making it hard to apply when dealing with multi-subject scenarios. In this study, we propose a highly-consistent data synthesis pipeline to tackle this challenge. This pipeline harnesses the intrinsic in-context generation capabilities of diffusion transformers and generates high-consistency multi-subject paired data. Additionally, we introduce **UNO**, which consists of progressive cross-modal alignment and universal rotary position embedding. It is a multi-image conditioned subject-to-image model iteratively trained from a text-to-image model. Extensive experiments show that our method can achieve high consistency while ensuring controllability in both single-subject and multi-subject driven generation. Code and model: https://github.com/bytedance/UNO.*

1. Introduction

As the material medium for abstract linguistic semantics and spatial embodiments of concrete visual subjects, images constitute the foundational modality in intelligent content generation. In recent years, customized image generation, which aims to create images that align with both the text semantics and the subjects in the reference images, has garnered significant interest across academic and industrial communities. This task unifies the flexibility of text control and the accuracy of visual controls, providing foundational infrastructure for diverse real-world applications ranging from film production to industrial design. As the field advances, the research challenge in customized image generation now centers on developing *a stable and scalable paradigm for unlocking more controllability*, *i.e.*, continuously increasing the amount of visual subject control without compromising the original text controllability.

Data, while serving as the foundation for training generative models, has long been the bottleneck in customized generation. An ideal model should be capable of generating visual subjects in diverse poses, locations, sizes, and other attributes based on text prompts. This necessitates data that encompasses multi-perspective subject variations, a requirement hindered by the impracticality of acquiring such comprehensive *real paired datasets*.

∗Corresponding author
† Project lead

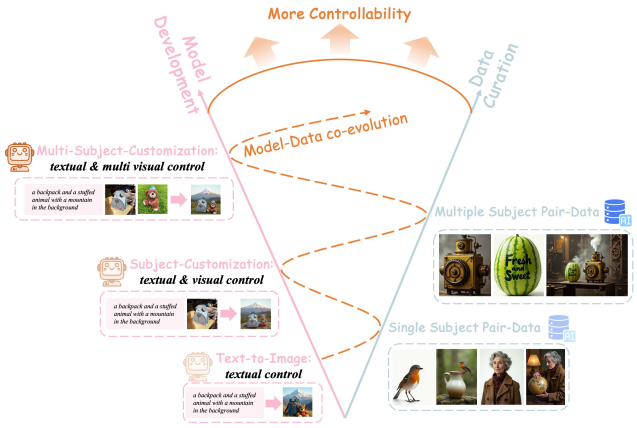

Figure 2. The illustration of our motivation. We propose a novel ***model-data co-evolution*** paradigm, where less-controllable preceding models systematically synthesize better customization data for successive more-controllable variants, enabling persistent co-evolution between enhanced model and enriched data.

Existing customized generation methods can be categorized into two streams based on *how they utilize the data to design the model*, *i.e.*, few-data fine-tuning and large-data training stream. Early few-data fine-tuning approaches [8, 33] primarily employ per-subject optimization through model fine-tuning [16, 33] or textual inversion [8], incurring substantial computational overhead and time-consuming during inference, which hinders real-world deployment. Therefore, more recent researches focus on the latter stream, which train adapters or image encoders on a large set of visual subjects to achieve real-time customization. Their corresponding data used for training are either real images with diversity limitations (eg, restricted subject variations [45, 47]), or synthetic data with limited image quality (typically, $\leq 512 \times 512$) and narrow domain coverage. Therefore, these methods often exhibit a trade-off between subject similarity and text controllability [45], or unstable generation [14]. Essentially, the existing customized models are designed based on their corresponding available customized data, resulting in limited scalability due to their data bottleneck.

Diverging from the existing *data-driven model design*, research on large language models (LLMs) demonstrates their capacity for strategic synthetic data generation toward model self-enhancement. Their bidirectional data-model knowledge transfer manifests through either high-performance models providing supervisory signals for weaker counterparts[1, 9], or conversely, less capable models could provide supervision to elicit higher capabilities leading to a stronger variant[3, 23, 35]. Inspired by