TWEEDIEMIX: IMPROVING MULTI-CONCEPT FUSION FOR DIFFUSION-BASED IMAGE/VIDEO GENERATION

Gihyun Kwon
KRAFTON
gkwon@krafton.com

Jong Chul Ye
Kim Jaechul Graduate School of AI, KAIST
jong.ye@kaist.ac.kr

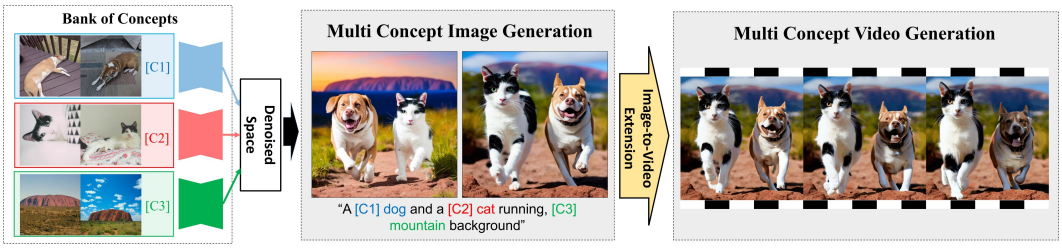

Figure 1: **Multi-concept Generation Results from TweedieMix.** Our model can generate high-quality multi-concept generation results on both of image and video domains. More results can be found in the experiment section.

ABSTRACT

Despite significant advancements in customizing text-to-image and video generation models, generating images and videos that effectively integrate multiple personalized concepts remains a challenging task. To address this, we present TweedieMix, a novel method for composing customized diffusion models during the inference phase. By analyzing the properties of reverse diffusion sampling, our approach divides the sampling process into two stages. During the initial steps, we apply a multiple object-aware sampling technique to ensure the inclusion of the desired target objects. In the later steps, we blend the appearances of the custom concepts in the de-noised image space using Tweedie's formula. Our results demonstrate that TweedieMix can generate multiple personalized concepts with higher fidelity than existing methods. Moreover, our framework can be effortlessly extended to image-to-video diffusion models, enabling the generation of videos that feature multiple personalized concepts. Results and source code are in our anonymous project page.[1]

1 INTRODUCTION

In recent years, text-to-image generation models (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022) have made remarkable strides, empowering creatives to produce high-quality images simply by crafting text prompts. This success has quickly expanded into other domains like video and 3D scene generation (Zhang et al., 2023a; Esser et al., 2023; Poole et al., 2022; Xu et al., 2023; Zhou et al., 2023; Liu et al., 2023a), achieving impressive results. Significant progress has been made in developing models that can customize images for specific subjects or visual concepts (Kumari et al., 2023; Gal et al., 2022; Ruiz et al., 2023; Tewel et al., 2023). These have enabled new possibilities for content creation, allowing users to leverage their own personalized characters.

Despite significant advancements in customizing these models for specific subjects or visual concepts, a major challenge persists: generating images that effectively combine multiple personalized concepts. Existing methods (Kumari et al., 2023; Tewel et al., 2023) allow for joint training of models on multiple concepts or merging customized models to create scenes featuring more than

[1] https://github.com/KwonGihyun/TweedieMix

one personalized element. However, these approaches often struggle with semantically related concepts—such as cats and dogs—and have difficulty scaling beyond three concepts. For instance, Mix-of-Show (Gu et al., 2023) attempted to tackle multi-concept generation using disentangled Low-Rank (LoRA) (Hu et al., 2021) weight merging and regional guidance during sampling. Yet, issues like concept blending remain due to the complexities involved in weight merging. To address these limitations, ConceptWeaver (Kwon et al., 2024) introduced a training-free method that combines multiple concepts during inference by splitting the generation process into multiple stages.

In this paper, we introduce an enhanced, tuning-free approach for composing customized text-to-image diffusion models during the inference stage. Unlike previous methods that require weight merging or additional inversion steps for multi-object generation, our technique utilizes only the reverse sampling steps and divides the process into two main stages. First, we conduct multi-object-aware sampling using text prompts that include multiple objects, introducing a novel resampling strategy to further improve generation quality. In the second stage, we integrate custom concept models through object-wise region guidance. To ensure stable and high-quality sampling, we combine each custom concept sample within the intermediate denoised image space calculated using Tweedie's formula. To expand the versatility of our method, we also propose a training-free strategy for extending these custom concept-aware images into the video domain.

Our experimental results demonstrate that our method can compose images featuring semantically related concepts without incorrectly blending their appearances. Moreover, our model seamlessly handles more than two concepts, overcoming a common limitation of baseline approaches. The images generated closely align with the semantic intent of the input prompts, achieving high CLIP scores. Finally, our video outputs outperform existing fine-tuning-based custom video generation methods, underscoring the effectiveness of our proposed framework.

2 Related Work

Text-to-image (T2I) generation models have seen remarkable advancements over the years, evolving from early Generative Adversarial Network (GAN)-based models (Esser et al., 2021; Zhang et al., 2017) to the latest diffusion-based approaches (Saharia et al., 2022; Rombach et al., 2022; Yu et al., 2023; Ramesh et al., 2022). The development of Diffusion models has opened up a range of applications, including text-guided image editing (Hertz et al., 2023; Couairon et al., 2023; Mokady et al., 2022), image translation (Kwon & Ye, 2023; Tumanyan et al., 2023), and style transfer (Zhang et al., 2023c). Recently, the success of T2I models has seamlessly extended to other modalities, such as 3D scene and asset generation (Poole et al., 2022; Xu et al., 2023), as well as video generation (Zhang et al., 2023a; Esser et al., 2023; Zhou et al., 2023; Bar-Tal et al., 2024). This expansion has spurred research into applications like Image-to-3D (Liu et al., 2023a; 2024) and Image-to-Video generation (Xing et al., 2023; Zhang et al., 2023b), along with editing capabilities for 3D scenes (Park et al., 2024; Chen et al., 2023) and videos (Jeong et al., 2023; Ceylan et al., 2023).

Building upon these advancements, there has been a growing interest in customizing T2I models using user-provided images or visual concepts. The pioneering work of Textual Inversion (Gal et al., 2022) focused on optimizing textual embeddings to represent custom concepts, enabling the generation of images that reflect these custom concepts. Subsequent studies have enhanced performance by developing extended textual embeddings (Voynov et al., 2023; Li et al., 2023) and fine-tuning model parameters (Kumari et al., 2023; Ruiz et al., 2023; Tewel et al., 2023), leading to more efficient and flexible customization options.

Extending beyond single-concept frameworks, researchers have explored methods for incorporating multiple concepts into customized models which use joint training to embed multiple concepts simultaneously or weight merging of single-concept customized model parameters (Kumari et al., 2023; Han et al., 2023; Tewel et al., 2023). However, these methods face challenges when scaling to a larger number of concepts or when dealing with semantically similar concepts, often resulting in the concept blending or disappearance of specific concepts. To address these issues, recent work like Mix-of-Show (Gu et al., 2023) applied regional guidance during the sampling process using merged weights to mitigate concept blending. Despite this improvement, the approach still requires additional optimization steps for weight merging.