Many-to-many Image Generation with Auto-regressive Diffusion Models

Ying Shen[◊], Yizhe Zhang[†], Shuangfei Zhai[†], Lifu Huang[◊], Josh Susskind[†], Jiatao Gu[†]

[†]Apple [◊]Virginia Tech

Abstract

Recent advancements in image generation have made significant progress, yet existing models present limitations in perceiving and generating an arbitrary number of interrelated images within a broad context. This limitation becomes increasingly critical as the demand for multi-image scenarios, such as multi-view images and visual narratives, grows with the expansion of multimedia platforms.

This work introduces a domain-general framework for many-to-many image generation, capable of producing interrelated image series from a given set of images, offering a scalable solution that obviates the need for task-specific solutions across different multi-image scenarios.

To facilitate this, we present MIS, a novel large-scale multi-image dataset, containing 12M synthetic multi-image samples, each with 25 interconnected images. Leveraging MIS, we propose a domain-general Many-to-many Diffusion (M2M) model, a conditional diffusion model that can perceive and generate an arbitrary number of interrelated images in an auto-regressive manner, thus offering the flexibility and adaptability needed to meet a broad range of multi-image generation tasks.

Multi-Image Set Dataset (MIS)

We introduce MIS, the first large-scale multi-image dataset comprising sets of images interconnected by general semantic relationships. MIS consists of 12M synthetic multi-image set samples, each containing 25 interconnected images. Designed for broad, domain-general multi-image generation.

Specifically, we leverage the power of the Latent Diffusion Model and its capacity to generate a diverse set of images from the same caption by employing different latent noises, ensuring coherence and uniqueness within each set.

#jellyfish #blue #ocean #pretty SeaTurtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn And <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life

Many-to-many Diffusion (M2M)

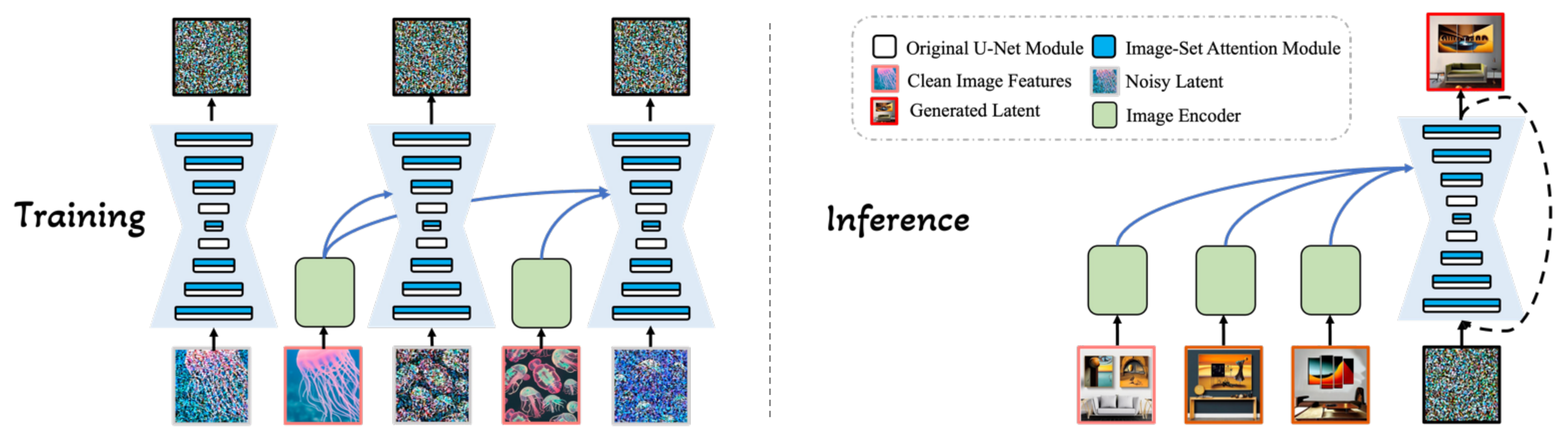
We introduce the Many-to-many Diffusion (M2M) framework, designed to perceive and generate an arbitrary number of interrelated images auto-regressively.

M2M adpats the pre-trained Stable Diffusion by replacing the text-to-image cross-attention module with our Image-Set Attention module. This allows the model to learn and understand the intricate interconnections within a set of images, thereby facilitating contextual coherence in multi-image generation.

M2M explores various architectural approaches for multi-image generation, with a focus on how preceding images are encoded. We discuss two main model variants: the **M2M with Self-encoder (M2M-Self)** and the **M2M with DINO encoder (M2M-DINO)**.

M2M-Self leverages the U-Net-based denoising model to simultaneously process the preceding and the noisy latent images, enabling cross-attention mechanisms over various spatial dimensions of the preceding images.

M2M-DINO explores integrating external vision models to encode preceding images, aiming to complement the U-Net's inherent capabilities for encoding images.



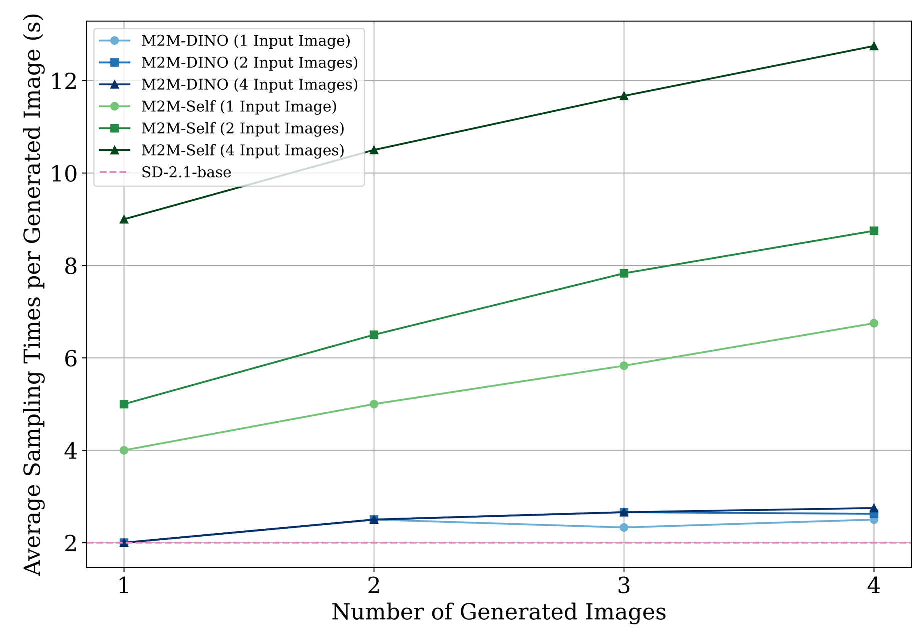
Quantitative Results

Quantitative Evaluation on 10K MIS Test Subset. Each metric is reported as an average score ± standard deviation across the 10 generated images.

Method	FID ↓	IS ↑	Text-Image CLIP ↑	Image-Image CLIP ↑
M2M-Self (9M)	9.56 ± 1.21	26.19 ± 0.67	22.71 ± 0.52	76.29 ± 0.02
M2M-DINO (6M)	8.88 ± 0.87	28.07 ± 0.58	23.05 ± 0.49	77.41 ± 0.03

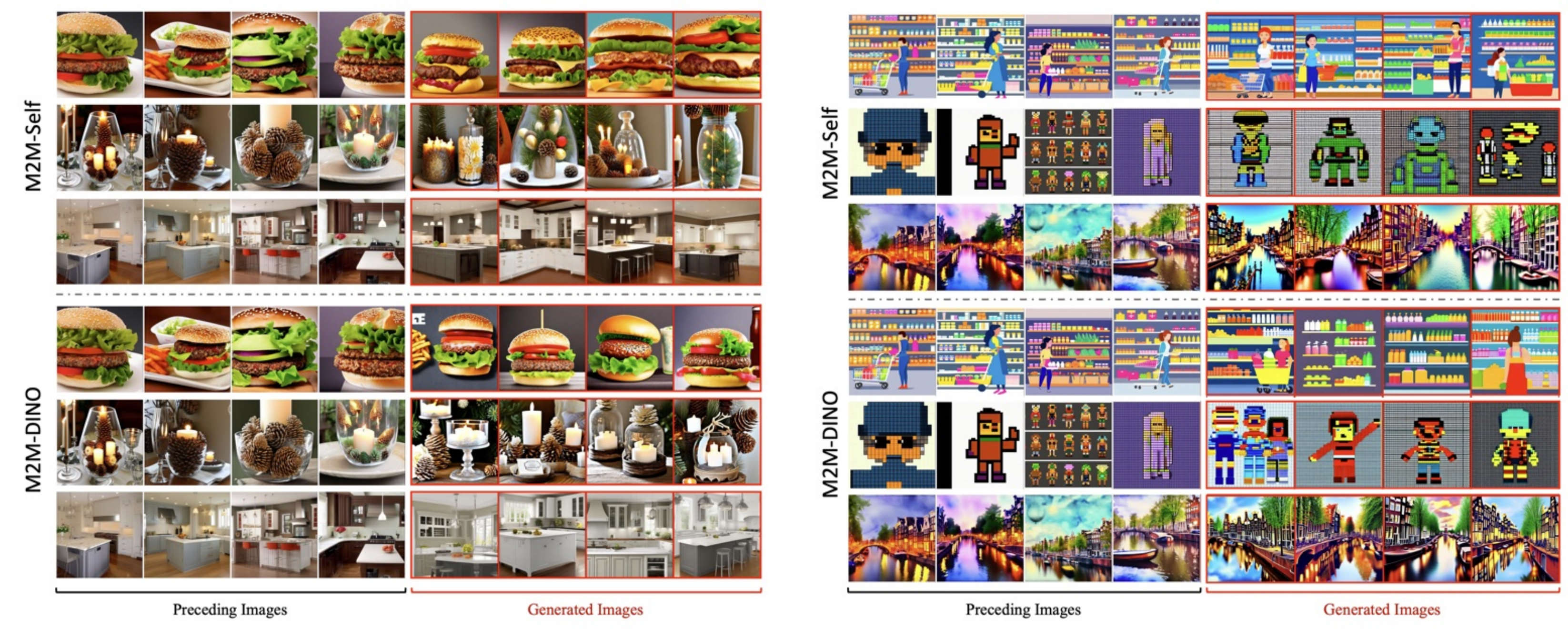
Sampling Efficiency

The sampling speed is measured as the average time to generate one image when using the DDIM sampler with 50 denoising steps. The efficiency is measured across M2M-Self and M2M-DINO when using 1, 2, and 4 input images, and compared against the StableDiffusion-2.1-base.



Ability to Capture the Relationship/Patterns

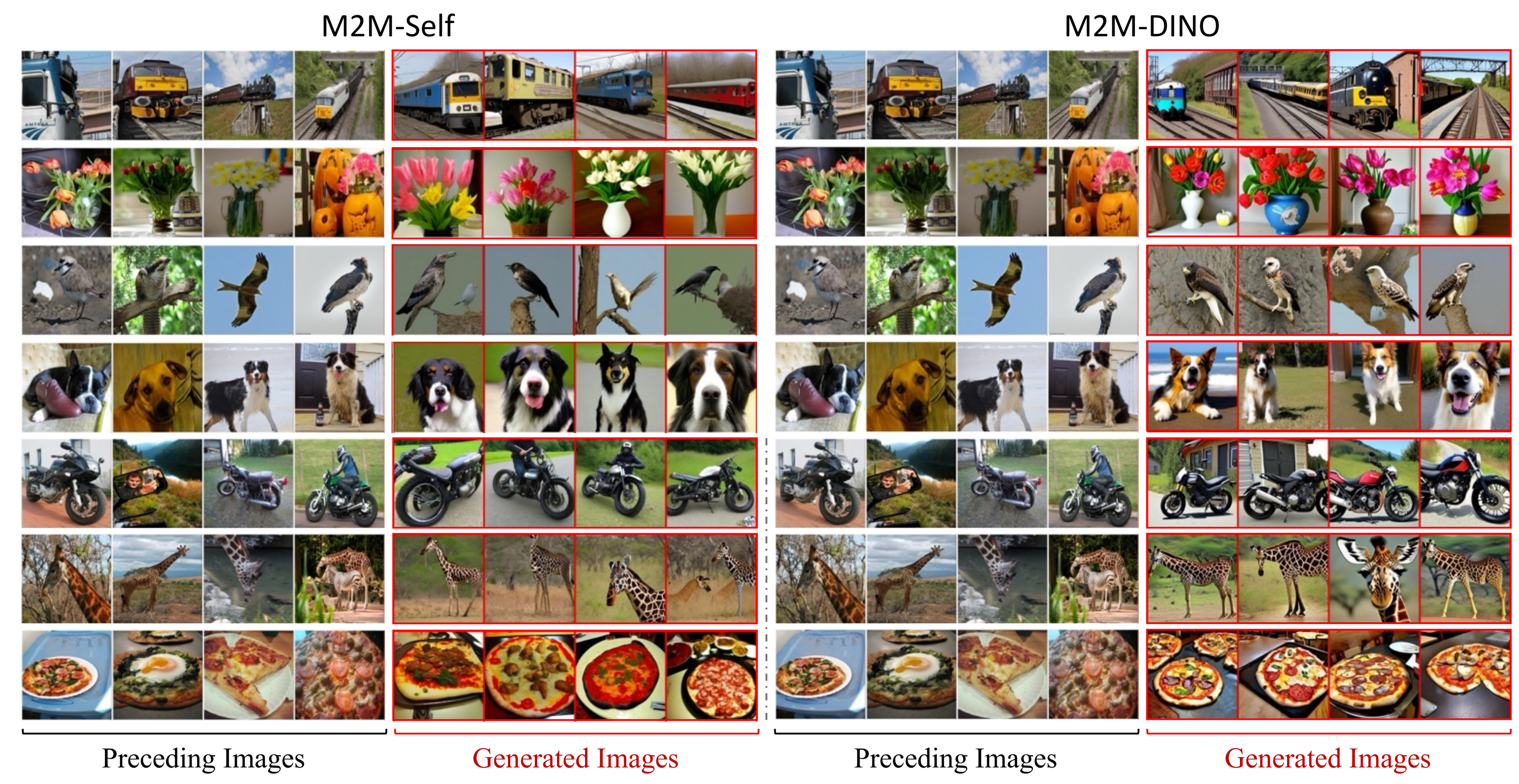
M2M captures style and content from preceding images and generates novel images in alignment with the observed patterns.



(a) Content Consistency (b) Style Consistency

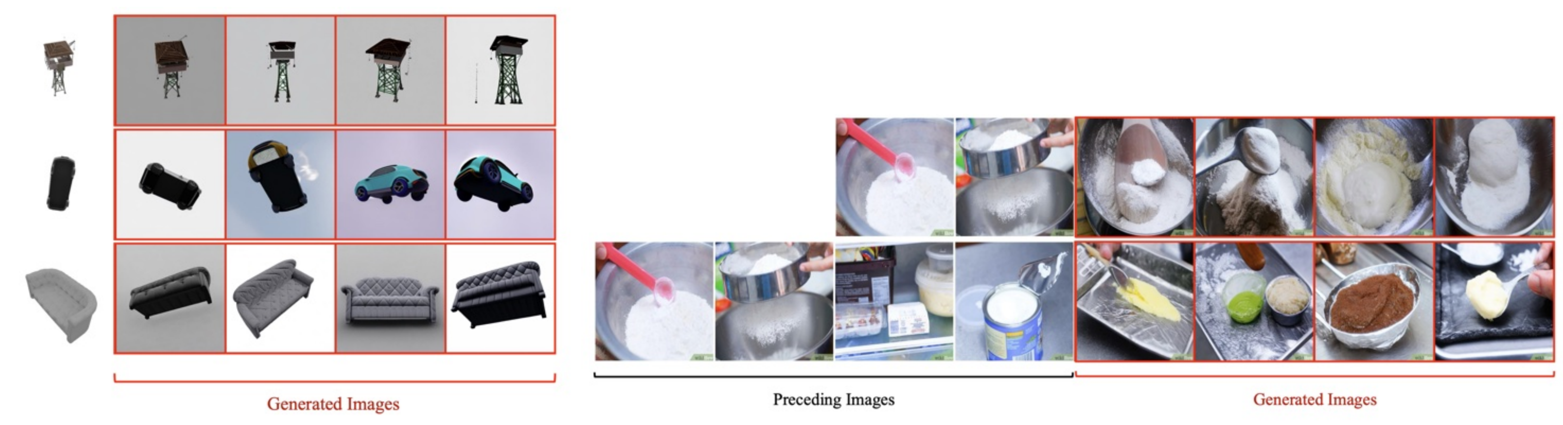
Generalization to Real Images

Impressively, despite being trained solely on synthetic data, M2M also exhibits zero-shot generalization to *real* images.



Adaptation for Specific Multi-Image Tasks

Building upon the initial training on MIS, we extend M2M's capabilities through task-specific fine-tuning for two different multi-image generation tasks: Novel View Synthesis and Visual Procedure Generation.



(a) Novel View Synthesis (b) Visual Procedure Generation

More Qualitative Results

Paper

Please check out our paper for more details!