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

Ying Shen [1]   **Yizhe Zhang** [2]   **Shuangfei Zhai** [2]   **Lifu Huang** [1]   **Josh Susskind** [2]   **Jiatao Gu** [2]

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 paper 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. Utilizing Stable Diffusion with varied latent noises, our method produces a set of interconnected images from a single caption. Leveraging MIS, we learn Many-to-many Diffusion (M2M), an autoregressive model for many-to-many generation, where each image is modeled within a diffusion framework. Throughout training on the synthetic MIS, the model excels in capturing style and content from preceding images — synthetic or real — and generates novel images following the captured patterns. Furthermore, through task-specific fine-tuning, our model demonstrates its adaptability to specific multi-image generation tasks, like Visual Procedure Generation.

1. Introduction

Over the past few years, the realm of image generation has seen remarkable progress across numerous tasks, including
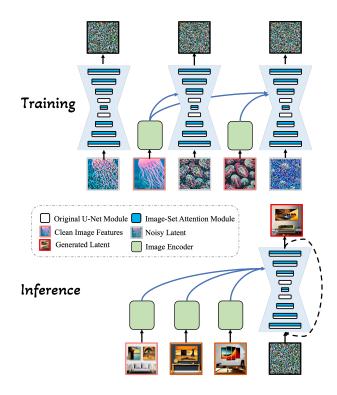

Figure 1: **Overview of M2M Pipeline.** During training, M2M takes sets of noised latent images and their corresponding clean image features as inputs and predicts the noise added to each noised latent image conditioned on the previous clean image features. At inference, our model can generate an arbitrary number of images in an auto-regressive manner by iteratively incorporating generated images back into the input.

super-resolution (Ho et al., 2022b; Saharia et al., 2022b), image manipulation (Meng et al., 2021; Nichol et al., 2022; Kawar et al., 2023; Brooks et al., 2023), text-to-image generation (Ramesh et al., 2021; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022a; Yu et al., 2022), and more. Despite this progress, the primary focus has predominantly been on the generation or processing of individual images. Some studies (Li et al., 2019; Maharana et al., 2022; Pan et al., 2022; Liu et al., 2023a) explore beyond single image generation but usually with a specific focus on specific tasks like story synthesis (Li et al., 2019), visual in-context learning (Bar et al., 2022), novel-view synthesis (Liu et al.,

---
[*]Equal contribution [1]Department of Computer Science, Virginia Tech, USA [2]Apple, USA. Correspondence to: Ying Shen <yings@vt.edu>, Jiatao Gu <jgu32@apple.com>.

2023c), and so on. These examples illustrate the versatility of the multi-image generation paradigm, which can also encompass video generation, essentially a sequence of image frames. The proliferation of multimedia platforms has increased the need for accepting multiple images and generating multiple images that possess different types of interconnection, catering to applications like producing thematically and stylistically consistent image sets for advertising or displaying objects from various perspectives. Recognizing the merging demand, this paper thus underscores the need for a holistic exploration into the general-domain multi-image to multi-image generation paradigm, where models are designed to perceive and generate an arbitrary number of interrelated images within a broader context.

In this work, we present a domain-general framework for multi-image to multi-image generation that can perceive and generate a flexible number of interrelated images auto-regressively, thus offering the flexibility and adaptability needed to meet a broad range of multi-image generation tasks. A cornerstone of this endeavor is the exposure of our framework to a diverse collection of multi-image examples that inherently maintain meaningful interrelations amongst the images in each set. To facilitate this, we introduce MIS, the first large-scale multi-image dataset comprising sets of images interconnected by general semantic relationships. Unlike previous multi-image datasets specialized towards specific scenarios, such as sequential frames or images from multiple viewpoints, MIS encapsulates more general semantic interconnections among images. MIS consists of a total of 12M synthetic multi-image set samples, each containing 25 interconnected images. Motivated by the success of diffusion models in text-to-image generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022a; Nichol et al., 2022), we propose to utilize Stable Diffusion to produce interconnected image sets from identical caption but varied latent noise. This ensures that all images generated from the same caption are drawn from the same conditional probability distribution, underscoring their meaningful internal connections. Yet, the varied latent noise guarantees the distinctiveness of each image within the set.

Leveraging our MIS dataset, we propose Many-to-many Diffusion (M2M), a conditional diffusion model that can perceive and generate an arbitrary number of interrelated images in an auto-regressive manner, based purely on visual inputs, without reliance on textual descriptions. M2M is built upon latent diffusion models and we extend it into a multi-image to multi-image generator by introducing an Image-Set Attention module that learns to capture the intricate interconnections within a set of images, thereby facilitating more contextually coherent multi-image generation. Our study further explores various architectural designs for multi-image to multi-image generation, focusing on diverse strategies for handling preceding images. We introduce

two novel model variants: M2M with Self-encoder (M2M-Self) and M2M with DINO encoder (M2M-DINO). M2M-Self utilizes the same U-Net-based denoising model to simultaneously process preceding latent images along with noisy latent images, enabling more refined cross-attention. Meanwhile, M2M-DINO employs external vision models to encode more discriminative visual features from preceding images. Experimental results show that M2M learns to capture style and content from preceding images and generate novel images in alignment with the observed patterns. Impressively, despite being trained solely on synthetic data, M2M exhibits zero-shot generalization to *real* images. Furthermore, through task-specific fine-tuning, our model demonstrates its adaptability to specific multi-image generation tasks, like Visual Procedure Generation, suggesting its potential to handle complex multi-image generation challenges.

Our paper makes the following contributions: (1) We introduce an innovative strategy for constructing MIS, the first large-scale multi-image dataset containing 12M synthetic multi-image set samples, each with 25 images interconnected by general semantic relationships. (2) We propose a domain-general Many-to-many Diffusion model that can perceive and auto-regressively generate an arbitrary number of interrelated images. (3) We show that M2M captures style and content from preceding images to generate novel images following the captured patterns. It exhibits great zero-shot generalization to real images and notable potential for customization to specific multi-image generation tasks.

2. Related Work

2.1. Image Generation

Image generation has always been a heated topic in computer vision, with demand spanning various tasks, such as super-resolution (Ho et al., 2022b; Saharia et al., 2022b), image manipulation (Meng et al., 2021; Nichol et al., 2022; Kawar et al., 2023; Brooks et al., 2023), text-to-image generation (Ramesh et al., 2021; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022a; Yu et al., 2022), and so on. Beyond the realm of single-image generation, there has been an exploration into multi-image generation (Li et al., 2019; Bar et al., 2022; Liu et al., 2023c) but usually with a specific focus on specific tasks. For instance, story synthesis (Li et al., 2019) aims at generating a series of images that narrate a coherent narrative. Other tasks, such as visual in-context learning (Bar et al., 2022) focus on generating one target image using a query image accompanied by example image pairs. Additionally, novel-view synthesis (Liu et al., 2023c) aims to generate images from new viewpoints based on a set of posed images of a particular scene or object. Our work distinctively expands on the existing literature by proposing a more general multi-image to multi-image

generation framework, that can be adapted to a myriad of scenarios in multi-image generation.

2.2. Diffusion-based Generative Models

Diffusion-based generative models(Ho et al., 2020; Song & Ermon, 2019) have marked notable progress in the field of generative AI. These models operate by incrementally introducing noise to perturb the data and learning to generate data samples from Gaussian noise through a series of denoising processes. Recently, diffusion models have become prominent for generating images (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022a; Brooks et al., 2023), as well as in other modalities such as video (Ho et al., 2022a; Singer et al., 2022; Blattmann et al., 2023), audio (Liu et al., 2023b; Huang et al., 2023), text (Li et al., 2022; Gong et al., 2022; Zhang et al., 2023), 3D (Poole et al., 2022; Gu et al., 2023; Liu et al., 2023c) and more. Given their superior performance in image generation, we propose to leverage the diffusion-based generative models for two key purposes: constructing the synthetic multi-image set dataset, MIS, and serving as the backbone architecture for our multi-image generation framework.

3. Background

The Diffusion Probabilistic Models (Ho et al., 2020; Song & Ermon, 2019) learns the data distribution $p(x)$ by gradually denoising a normally distributed variable throughout a Markov chain with length T. Specifically, diffusion models define a forward diffusion process in a Markov chain, incrementally adding Gaussian noise samples to the initial data point x into the Gaussian noise x_T over T steps, and a learnable reverse process that denoises x_T back to the clean input x iteratively via a sequence of time-conditioned denoising autoencoders $\epsilon_\theta(x_t, t)$. Typically, the denoising model ϵ_θ is implemented via time-conditioned U-Net (Ronneberger et al., 2015). The diffusion model is commonly trained with a simplified L2 denoising loss (Ho et al., 2020):

$$\mathcal{L}_{DM} = \mathbb{E}_{x,\epsilon,t}\big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\big], \qquad (1)$$

where $\epsilon \sim \mathcal{N}(0, 1)$, $t \sim \mathcal{U}(0, T)$.

Latent Diffusion Models Latent Diffusion Models (Rombach et al., 2022) improves the efficiency of diffusion models by operating in the latent representation space of a pretrained variational autoencoder (Kingma & Welling, 2013) with encoder $\mathcal{E}$ and decoder $\mathcal{D}$, such that $\mathcal{D}(\mathcal{E}(x)) \approx x$.

Diffusion models can be conditioned on various signals such as class labels or texts. The conditional latent diffusion models learn a denoising model ϵ_θ that predicts the noise added to the noisy latent z_t given conditioning c via the

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

Figure 2: A sample image set of five distinctive images generated using a caption from CC12M.

following objective:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),c,\epsilon,t}\big[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\big]. \qquad (2)$$

During inference, classifier-free guidance (Ho & Salimans, 2021) is used to improve sample quality:

$$\tilde{\epsilon}_\theta(z_t, c) = \epsilon_\theta(z_t, \varnothing) + s \cdot (\epsilon_\theta(z_t, c) - \epsilon_\theta(z_t, \varnothing)), \qquad (3)$$

where $\epsilon_\theta(z_t, c)$ and $\epsilon_\theta(z_t, \varnothing)$ refer to the condition and unconditional ϵ-predictions, and s represents the guidance scale. Setting $s = 1$ disables the classifier-free guidance while increasing $s > 1$ strengthens the effect of guidance.

4. Multi-Image Set Construction

Learning to process and generate multiple images requires a diverse collection of multi-image examples for training. As manually collecting a multi-image dataset comprising sets of interconnected images would be resource-intensive, we propose to leverage the capabilities of text-to-image models (Latent Diffusion Model (Rombach et al., 2022)) for generating a multi-image dataset, which we refer to as MIS. This dataset comprises sets of interconnected images. 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. This approach allows us to construct a set of interconnected images for each caption, recognizing that these images within the set exhibit internal connections owing to the common caption they were generated with. Notably, while all images generated from the same caption are drawn from the same conditional probability distribution, underscoring their meaningful internal connections, the introduction of varied latent noise ensures the distinctiveness of each image within the set.

To gather image captions, we employ Conceptual 12M (Changpinyo et al., 2021), a large image-text pair dataset that contains approximately 12 million web images, each with a corresponding descriptive alt-text. For MIS generation, we exclusively utilize these alt-texts as captions for generating interconnected images. More precisely, with each caption, we employ the Stable Diffusion model to generate a set of distinct images by using different latent

noise. We employ Stable Diffusion v2-1-base [1] along with the Euler Discrete Scheduler (Karras et al., 2022) and a default guidance scale of 7.5 for image generation. Stable Diffusion generates images conditioned on CLIP (Radford et al., 2021) text embeddings. To comply with the maximum token length allowed for the CLIP Text encoder, we filter out captions exceeding the maximum token length of 77 tokens. This results in a curated collection of 12,237,187 captions for our multi-image generation. For each individual caption, we employ different latent noises to generate 25 distinct images. As a result, our final MIS consists of 12M multi-image set samples, with each image set containing 25 interconnected images. Figure 2 shows an example of an image set generated using a single caption from Conceptual 12M.

5. 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, as shown in Figure 1. Our framework adpats the pre-trained Stable Diffusion, a text-to-image latent diffusion model, which uses U-Net as the denoising model $\epsilon_\theta(\cdot)$, 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)** in Section 5.1 and the **M2M with DINO encoder (M2M-DINO)** in Section 5.2. 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. Meanwhile, M2M-DINO explores integrating external vision models to encode preceding images, aiming to complement the U-Net's inherent capabilities for encoding images.

5.1. M2M with Self-encoder (M2M-Self)

During training, given an image set $\{I\}_{i=1}^N$, each image I_i is encoded individually into the latent code z_0^i using a pre-trained autoencoder $z_0^i = \mathcal{E}(I_i)$. These image latents are stacked to construct $z_0^{1:N} \in \mathbb{R}^{N \times C \times H \times W}$, where N represents the number of images within each set, C is the number of latent channels, and H and W are the spatial dimensions of the latent space. The clean latent $z_0^{1:N}$ is then noised according to the pre-defined forward diffusion schedule to produce the noisy latent $z_t^{1:N}$, where the noise

---
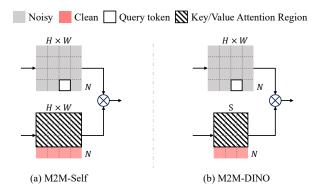[1]https://huggingface.co/stabilityai/stable-diffusion-2-1-base

Figure 3: **Illustration of Image-Set Attention Module.** The query token is denoted in a white square and its corresponding key/value attention region is marked by a diagonal striped pattern.

level increases over diffusion timesteps t.

To prepare the input for M2M-Self, we concatenate the clean and noisy latent codes, resulting in $z^{1:N} = [z_0^{1:N}; z_t^{1:N}] \in \mathbb{R}^{2 \times N \times C \times H \times W}$. Here, the symbol ; denotes the concatenation operation and the factor 2 indicates the inclusion of both clean and noised latent forms. M2M-Self thus processes the concatenated tensor $z^{1:N}$ and predicts the noise added to each noised latent image auto-regressively.

Image-Set Attention The core component in M2M-Self is the Image-Set Attention module, designed for effective cross-attention from noisy latent to their preceding clean ones. The module accepts an input tensor $z \in \mathbb{R}^{BZ \times 2 \times N \times H \times W \times C}$, where BZ is the batch size. For readability, the superscript $1 : N$ has been omitted. The input z is first partitioned and reshaped into two components: the noisy latent $z_n \in \mathbb{R}^{BZ \times (N \times H \times W) \times C}$ and its corresponding clean latent $z_c \in \mathbb{R}^{BZ \times (N \times H \times W) \times C}$. This transformation enables a comprehensive cross-attention across the resultant length of $N \times H \times W$. These two latents are then projected and processed via cross-attention (Vaswani et al., 2017):

$$z'_n = \text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V, \quad (4)$$

where $Q = W^Q z^Q$, $K = W^K z^K$, and $V = W^V z^V$, with d indicating latent feature dimension. The cross-attention operates with the noisy latents as queries $z^Q = z_n$ and the clean latents as keys and values $z^K = z^V = z_c$. We also employ multi-head cross-attention, allowing the model to attend to information from different representation subspace (Vaswani et al., 2017). To ensure that each noisy latent z_n^i only attends to patches within its preceding clean latents $z_c^{<i}$, a causal image-set attention mask is introduced. Figure 3(a) visualizes this process with an example of a query token and the key/value region it allows to attend.

The output latent z' is formed by concatenating the original clean latent and the updated noisy latent as $z' = [z_c; z'_n]$.

5.2. M2M with DINO encoder (M2M-DINO)

Alongside M2M-Self, which uses the same U-Net-based denoising model for processing the preceding images as internal features, M2M-DINO incorporates external vision models to enhance the encoding of preceding images with more discriminative visual features. Specifically, we use DINOv2 (Oquab et al., 2023), known for its superior performance in understanding fine-grained vision information.

Image-Set Attention In M2M-DINO, each image I_i from a set $\{I\}_{i=1}^N$ is encoded into the latent code z_0^i using a pre-trained autoencoder $z_0^i = \mathcal{E}(I_i)$ and DINO features $v^i = \mathcal{E}_v(I_i)$ with the DINO image encoder $\mathcal{E}_v(\cdot)$. These encoded forms are then stacked separately to form $z_0^{1:N} \in \mathbb{R}^{N \times C \times H \times W}$ and $v^{1:N} \in \mathbb{R}^{N \times S \times D}$, where N is the number of images, S the length of DINO tokens, and D the dimensionality of DINO features. Subsequently, noisy latents $z_t^{1:N}$ are constructed from $z_0^{1:N}$ following a predefined noise addition schedule. For clarity, the superscript $1:N$ will be omitted.

The Image-Set Attention module takes in the noisy latent $z_t \in \mathbb{R}^{BZ \times N \times H \times W \times C}$ and the DINO features of clean images $v \in \mathbb{R}^{BZ \times N \times S \times W}$. It reshapes z_t into $\mathbb{R}^{BZ \times (N \times H \times W) \times C}$ and v into $\mathbb{R}^{N \times (N \times S) \times D}$ to facilitate the cross-attention. DINO features v act as keys and values $z^K = z^V = v$ and the noisy image latents z_t as queries z^Q, enabling interaction between noisy and clean features as depicted in Figure 3(b). Like M2M-Self, M2M-DINO utilizes a causal image-set attention mask, permitting each noisy latent z_t^i to exclusively attend to its preceding DINO features $v^{<i}$.

6. Training and Inference Procedure

6.1. Initial Training and Inference

Training Objective The training process is similar to the Latent Diffusion Model (Rombach et al., 2022). We initialize the model weights from the Stable Diffusion Model, but intentionally exclude the pre-trained text-to-image cross-attention layer, and initialize our Image-Set Attention module from scratch. M2M takes the visual features of the clean images, $z_0^{1:N}$, and the noisy latent images, $z_t^{1:N}$, as inputs. The training objective is to predict the noise strength added to each noised latent image conditioned on the previous clean image features. We employ an auto-regressive training strategy, where the loss is accumulated over the difference in the predicted noise of each image:

$$\mathcal{L}_\theta = \mathbb{E}_{\mathcal{E}(x_0^{1:N}), \epsilon, t}\left[\|\epsilon - \epsilon_\theta(z_t^{1:N}, z_0^{1:N}, t)\|_2^2\right], \quad (5)$$

where $z_0^{1:N}$ denotes the clean image latents in M2M-Self and DINO features in M2M-DINO.

Inference As shown in Figure 1, M2M performs auto-regressive image generation during inference by iteratively incorporating the images generated in the previous iteration as new inputs for next iterations. This process begins by generating an initial image, given a set of conditional input images, and randomly sampled noise latent codes. Afterward, the model incorporates the previously generated image into the input context for the next generation cycle. In practice, the number of context images that can be input into the model is limited by the predefined context window W. At inference time, with a guidance scale $s \geq 1$, the output of the model is extrapolated further in the direction of the conditional $\epsilon_\theta(z_t, z_0^{1:W}, t)$ and away from the unconditional $\epsilon_\theta(z_t, \mathbf{0}, t)$:

$$\begin{aligned}
\tilde{\epsilon}_\theta(z_t, z_0^{1:W}, t) &= \epsilon_\theta(z_t, \mathbf{0}, t) \\
&+ s \cdot (\epsilon_\theta(z_t, z_0^{1:W}, t) - \epsilon_\theta(z_t, \mathbf{0}, t)).
\end{aligned} \quad (6)$$

Here, $z_0^{1:W}$ is the sequence of context images used as conditional inputs in the generative process.

6.2. Task-specific Fine-tuning and Inference

Fine-tuning Building upon the initial training on MIS, we extend M2M's capabilities through task-specific fine-tuning for specific multi-image generation task. This involves integrating task-specific conditions c_C, such as positional information, camera viewpoints, etc., through additional embeddings. For M2M-Self, conditions are incorporated into the hidden states before applying Image-Set Attention as $z^{1:N} = z^{1:N} + E(c_C)$, where $E(\cdot)$ is a general embedding strategy adapting to different conditions c_C. In M2M with DINO encoder, which uses DINO features for encoding preceding images, a distinct conditional embedder, $E_c(\cdot)$, is applied, represented by $v^{1:N} = v^{1:N} + E_c(c_C)$. This approach enables the model to incorporate additional contextual or spatial information relevant to the task, enhancing its capability to generate images aligned with the task-specific requirements. Similarly, to facilitate classifier-free guidance, we randomly zero out the task-specific conditions along with the clean image condition $z_0^{1:N}$ with 10% probability.

Inference During inference, M2M employs distinct guidance scales, s_I for the clean image condition $z_0^{1:W}$ and s_C for task-specific conditions c_C, enabling refined control over image generation based on both image context and task

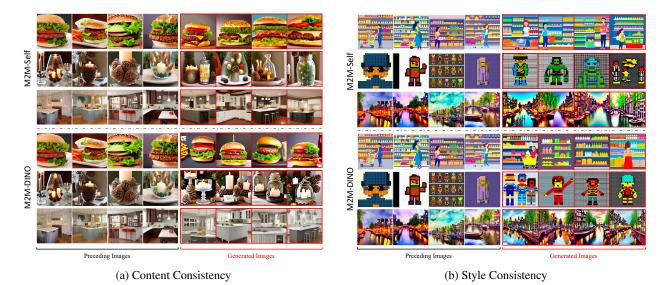(a) Content Consistency

(b) Style Consistency

Figure 4: **Consistency Evaluation in M2M-Self and M2M-DINO**: The figure shows the ability of M2M-Self and M2M-DINO to maintain content (a) and style (b) consistency. Content consistency refers to the model's capacity to generate images with the same type of subject as preceding ones, while style consistency pertains to maintaining aesthetic elements like color schemes, textures, and artistic techniques. Each subfigure contains two panels: the top for M2M-Self and the bottom for M2M-DINO. Columns 1-4 show the conditioning images, and Columns 5-8 display generated images.

needs. Consequently, Equation 6 is adapted as follows:

$$
\begin{aligned}
\tilde{\epsilon}_\theta&(z_t, z_0^{1:W}, c_C) \\
&= \epsilon_\theta(z_t, \varnothing, \varnothing) \\
&+ s_I \cdot \big(\epsilon_\theta(z_t, z_0^{1:W}, \varnothing) - \epsilon_\theta(z_t, \varnothing, \varnothing)\big) \\
&+ s_C \cdot \big(\epsilon_\theta(z_t z_0^{1:W}, c_C) - \epsilon_\theta(z_t, z_0^{1:W}, \varnothing)\big). \quad (7)
\end{aligned}
$$

7. Experimental Setup

7.1. Datasets

Pre-training Datasets   For pre-training, we leverage our introduced MIS, as detailed in Section 4. We adopt two distinct subsets of this dataset for training two model variants: M2M with Self-encoder (M2M-Self) and M2M with DINO encoder (M2M-DINO). Specifically, the M2M-Self model is trained on a subset of 9M multi-image examples, each containing a set of $N = 5$ images. Meanwhile, M2M-DINO is trained on a subset consisting of 6M multi-image examples.

Task-Specific Fine-tuning Datasets   Visual Goal-Step Inference (VGSI) (Yang et al., 2021) consists of around 53K wikiHow articles across various everyday tasks. Each article contains one or more different methods to achieve it, with each method including a series of specific steps accompanied by corresponding images. We employ this dataset to construct the visual procedure generation task where the model generates images depicting future steps given images from preceding steps.

7.2. Evaluation Metrics

To assess the performance of our models, we employ Fréchet Inception Distance (FID) (Heusel et al., 2017) and inception score (IS) (Salimans et al., 2016) to measure the quality of generated images. Additionally, we utilize the CLIP score (Radford et al., 2021; Hessel et al., 2021) to measure the alignment of the generated images with their preceding images. Our evaluation is conducted on a random selection of 10K samples from the test split of MIS. Furthermore, we present both qualitative evaluations from this test split and real-world images.

8. Results and Discussion

8.1. Ability to Capture the Relationship/Patterns

In this section, we investigate the model's capacity to capture the relationship or patterns within preceding images and subsequently generate new images in alignment with the observed patterns. Specifically, we conduct experiments evaluating the following two key aspects: (1) Content consistency and (2) Style consistency. Content consistency evaluates the model's ability to generate images featuring the same type of subject as in the preceding images. Style consistency, on the other hand, examines the model's capability to maintain the aesthetic or stylistic aspects of preceding images, including color schemes, textures, and artistic techniques.

Content Consistency   In assessing content consistency, we applied M2M to the test subset of MIS. As shown in
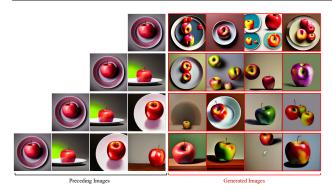
Figure 5: **Effect of Varying Preceding Images.** The figure presents the images generated from M2M-DINO when conditioned on varying numbers of preceding images.

Figure 4a, the images generated from M2M-Self (upper section) and M2M-DINO (lower section) demonstrate that the proficiency of both model variants in preserving content integrity across a varied range of subjects. From Row 1 to Row 3 for each model variant, we demonstrate that M2M adeptly maintains content consistency in images, starting from simple objects such as a hamburger, progressing to more complex compositions involving multiple objects (like a pinecone, candle, and glass), and extending to detailed indoor environments and scenes involving people.

Style Consistency   Style consistency is another critical aspect of our evaluation. As shown in Figure 4b, both M2M-Self and M2M-DINO effectively replicate a variety of artistic styles derived from preceding images. In each model panel, from Row 1 through Row 3, the model consistently maintains the styles of preceding images, ranging from simplified style to pixel art and watercolor paintings.

8.2. Effect of Number of Preceding Images

We further investigate the effect of altering the numbers of preceding images on multi-image generation. As illustrated in Figure 5, our experiment involves conditioning M2M-DINO with a different count of preceding images, specifically from one to four, each depicting a single apple. Initially, with only one preceding image as condition, M2M-DINO tends to generate images featuring an arbitrary number of apples, diverging from the single-apple pattern. However, as the number of preceding images increases, M2M-DINO begins to more accurately capture and replicate the pattern of a single apple. Our findings indicate that as the number of preceding images increases, the images generated by M2M-DINO are more likely to capture and reproduce the patterns observed in preceding images.

8.3. Generalization to Real Images

We evaluate the model's capability for zero-shot generalization to real images, which is pivotal in understanding
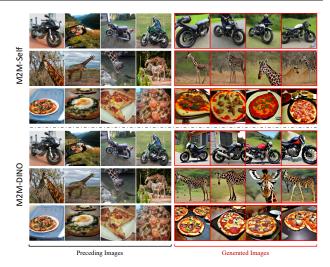

Figure 6: **Generalization to Real Images.** Columns 1-4 display the real images from MSCOCO, serving as the preceding images. Columns 5-8 show images generated based on the preceding images.

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

Table 1: **Quantitative Evaluation on 10K MIS Test Subset.** Each metric is reported as an average score ± standard deviation across the 10 generated images.

how well M2M can adapt to real-world scenarios beyond the synthetic data it was trained on. For this purpose, we employ MSCOCO (Lin et al., 2014), which contains real images of complex everyday scenes containing common objects. To assess the model's capability in maintaining content consistency across various real-world scenarios, we group images from MSCOCO into sets based on object categories. Each set contained different images, but all shared the same object category. Figure 6 showcases images generated (Columns 5-8) by M2M-Self and M2M-DINO, which are based on real images from the MSCOCO (Columns 1-4). Impressively, despite being trained solely on synthetic data, our model exhibits zero-shot generalization to *real* images. The generated images not only resemble the real images but also maintain a high degree of content consistency.

8.4. Quantitative Evaluation

In this section, we present a comprehensive quantitative assessment of M2M, focusing on the quality of the images generated by the model and their visual consistency with preceding images. Our evaluation leverages three established metrics: Fréchet Inception Distance (FID), Inception Score (IS), and various CLIP scores, utilizing a random subset of 10,000 samples from the MIS test split.

M2M is designed to accept a sequence of preceding images and autoregressively generate subsequent images. For this analysis, we generated 10 images using both M2M-Self and M2M-DINO, conditioned on four preceding images. The FID, IS, and CLIP scores were computed for every n-th image generated. FID is calculated by measuring the Fréchet distance between the multivariate Gaussian distributions of the 'real' (the first preceding images) and the generated images. To assess the consistency of the generated images with the preceding ones, we employ two variants of CLIP scores: text-image and image-image. The Text-Image CLIP score is calculated by comparing each n-th generated image to the common textual description associated with the preceding images in MIS. The Image-Image CLIP score measured the visual similarity of each n-th generated image with all preceding images. Table 1 details the average scores and standard deviations for these metrics, reported as an average score $\pm$ standard deviation across 10 generated images. Notably, M2M-DINO outperforms M2M-Self across all metrics, indicating a more robust capability in generating high-quality and contextually consistent images within a sequence.

8.5. Sampling Efficiency

We evaluate the sampling efficiency of our proposed M2M-DINO and M2M-Self methodologies against the Stable Diffusion-2.1-base (SD-2.1-base). Figure 7 illustrates the comparative analysis of sampling speed, measured as the average time required to generate a single image, across different model configurations, considering various numbers of input and generated images. All the models utilize the DDIM sampler with 50 denoising steps, and the evaluation is performed on a single NVIDIA A40 GPU to ensure a fair and consistent basis for comparison.

Our results indicate that M2M-DINO significantly outperforms M2M-Self in terms of sampling efficiency, especially as the number of input and generated images rises. Notably, M2M-DINO demonstrates a sampling speed on par with SD-2.1-base, even when M2M-DINO is set to process multiple input images and generate multiple images. These results highlight the robust and consistent efficiency of M2M-DINO in many-to-many image generation.

8.6. Adaptation for Specific Multi-Image Tasks

In this section, we present M2M fine-tuning results on Visual Procedure Generation, demonstrating its potential for customization to specific multi-image generation tasks. The Visual Procedure Generation task challenges M2M to understand and predict the sequence of visual steps in a procedure. To equip the model with the necessary understanding of sequence and progression, we introduce a positional embedder. This embedder utilizes sinusoidal encoding to capture
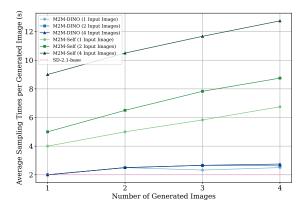

Figure 7: **Sampling Efficiency.** The sampling speed is measured as the average time to generate one image when using the DDIM sampler with 50 denoising steps on a single NVIDIA A40 GPU. The efficiency is measured across M2M-Self and M2M-DINO when using 1, 2, and 4 input images, and compared against the StableDiffusion-2.1-base.

Figure 8: **Visual Procedure Generation on VGSI** Columns 1-4: the sequence of historical visual steps. Columns 5 - 8: images generated by M2M-DINO that depict future visual steps.

the position of each image in the sequence, which is then processed through an MLP layer. Figure 8 showcases the model's ability to predict future steps in a visual procedure based solely on the sequence of input images.

9. Conclusion

We introduce MIS, a novel large-scale multi-image dataset, containing 12M synthetic multi-image samples, each with 25 interconnected images. We propose a domain-general Many-to-many Diffusion (M2M) model that can perceive and generate sequences of interrelated images autoregressively. We explore two main model variants, M2M-Self and M2M-DINO, both demonstrating strong ability in capturing and replicating style and content from preceding images when trained on MIS. Remarkably, our model exhibits zero-shot generalization to *real* images despite being trained solely on synthetic data. We further demonstrate the model's adaptability to specific multi-image generation tasks, like Visual Procedure Generation, through targeted fine-tuning, underscoring the potential of our approach to adapt to a broad spectrum of multi-image generation tasks.

References

Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., and Efros, A. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35: 25005–25017, 2022.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2022.

Gu, J., Trevithick, A., Lin, K.-E., Susskind, J. M., Theobalt, C., Liu, L., and Ramamoorthi, R. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pp. 11808–11826. PMLR, 2023.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022b.

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., and Zhao, Z. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/huang23i.html.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577, 2022.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6007–6017, 2023.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., and Gao, J. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6329–6338, 2019.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, C., Wu, H., Zhong, Y., Zhang, X., and Xie, W. Intelligent grimm–open-ended visual storytelling via latent diffusion models. *arXiv preprint arXiv:2306.00973*, 2023a.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. AudioLDM: Text-to-audio generation with latent diffusion models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/liu23f.html.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., and Vondrick, C. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023c.

Maharana, A., Hannan, D., and Bansal, M. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, pp. 70–87. Springer, 2022.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Pan, X., Qin, P., Li, Y., Xue, H., and Chen, W. Synthesizing coherent story with auto-regressive latent diffusion models. *arXiv preprint arXiv:2211.10950*, 2022.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022a.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022b.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yang, Y., Panagopoulou, A., Lyu, Q., Zhang, L., Yatskar, M., and Callison-Burch, C. Visual goal-step inference using wikihow. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2167–2179, 2021.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

Zhang, Y., Gu, J., Wu, Z., Zhai, S., Susskind, J., and Jaitly, N. Planner: Generating diversified paragraph via latent language diffusion model. *arXiv preprint arXiv:2306.02531*, 2023.

Appendix

A. Implementation Details

For our training procedure, we adopt two configurations for M2M with Self-encoder (M2M-Self) and M2M with DINO encoder (M2M-DINO). Both M2M-Self and M2M-DINO are trained with a total batch size of 256 on 8×80GB NVIDIA A100 GPUs for one epoch. We use a learning rate of 10^{-5} without any learning rate warm-up. M2M is initialized from the EMA weights of the Stable Diffusion v2-1 base [2]. For other configurations, we adopt the default training settings provided within the Stable Diffusion codebase. Regarding the encoding of preceding images in M2M with DINO encoder, we employ DINOv2-giant [3], particularly leveraging its last hidden states. To facilitate classifier-free guidance (Ho & Salimans, 2021), M2M is jointly trained with conditional and unconditional objectives. Training for unconditional denoising is achieved by randomly zeroing out the clean image condition $z_0^{1:N}$ with 10% probability. Similarly, when performing task-specific fine-tuning, we randomly zero out the task-specific conditions along with the clean image condition with 10% probability. At inference time, M2M generates novel images with 50 denoising steps using the DDIM sampler (Song et al., 2020). A guidance scale of 7.5 is employed for the preceding images unless specified otherwise.

B. Task-Specific Fine-tuning

We perform an additional task-specific fine-tuning task, Novel-View Synthesis. The Novel-View Synthesis task tests the model's capability to generate images from new viewpoints. This is achieved by integrating a camera embedding that encodes information about the camera's viewpoint. Specifically, this embedding captures the camera extrinsic for each image using a straightforward Multilayer Perceptron (MLP) layer. We finetune the model on Objaverse (Deitke et al., 2023), a large-scale dataset containing 800K+ 3D objects. Each object in the dataset contains 12 images of the object from different camera viewpoints along with the associated 12 camera poses. We utilize Objaverse to evaluate the model's performance in adapting to the novel view synthesis task. Given several posed images of a specific object, this task aims to generate images of the same object from novel viewpoints. Figure 9 demonstrates the images generated by M2M-DINO when conditioned on a singular image. We show that M2M-DINO is capable of auto-gressively generating multi-view images that are consistent with each other when conditioned on one initial preceding image.

---

[2]https://huggingface.co/stabilityai/stable-diffusion-2-1-base
[3]https://huggingface.co/facebook/dinov2-giant

Generated Images

Figure 9: **Novel View Synthesis on Objaverse** Column 1 presents a singular preceding image of an object. Columns 2 - 5 display the images generated by M2M-DINO from various novel viewpoints.

C. Limitations

While our model achieves considerable success in multi-image generation, it is not without its limitations. Notably, it struggles to generate human faces with high fidelity, a shortfall possibly stemming from the suboptimal quality of human faces present in our synthetic training set. Future efforts could benefit from incorporating more advanced diffusion models to enhance the quality of training data, particularly for human faces. Another observed challenge is the gradual decline in image quality during the auto-regressive generation of prolonged image sequences. This performance degradation highlights a potential area for further optimization, suggesting a need for improved strategies to maintain image quality throughout extended generative processes, which is critical for applications requiring the continuous production of images.

D. Impact Statement

We propose MIS, a novel large-scale multi-image dataset, containing 12M synthetic multi-image set samples, each with 25 interconnected images. This dataset is distinct from previous collections that focus on specific scenarios such as sequential frames or images from multiple viewpoints, by emphasizing a broader range of semantic interconnections. The creation of MIS is in response to the increasing demand on multimedia platforms for generating thematically and stylistically consistent image sets. Such capabilities are critical for applications ranging from digital advertising to virtual reality, where presenting objects from multiple perspectives or in varying thematic contexts is essential. Our research contributes a domain-general framework for multi-image to multi-image generation, capable of autoregressively perceiving and generating a variable number of interrelated images. This framework is designed to be flexible and adaptable, offering the potential for task-specific customization on various multi-image generation tasks.

E. Additional Experiments

E.1. Generation of Images Conditioned on Synthetic Images

Figures 10, 11, and 12 showcase images generated by the M2M-Self. On the other hand, Figures 13, 14, and 15 display the images generated from the M2M-DINO. These images were generated using the conditioning images from the test subset of the MIS dataset.

E.2. Generation of Images Conditioned on Real Images

Figure 16 and 17 present a comparative showcase of images generated by M2M-Self and M2M-DINO, respectively. These images are generated from real-world scenes contained in the MSCOCO dataset. The first four columns (Columns 1-4) display the original images from the MSCOCO dataset, whereas the generated images are presented in the subsequent columns (Columns 5-8).
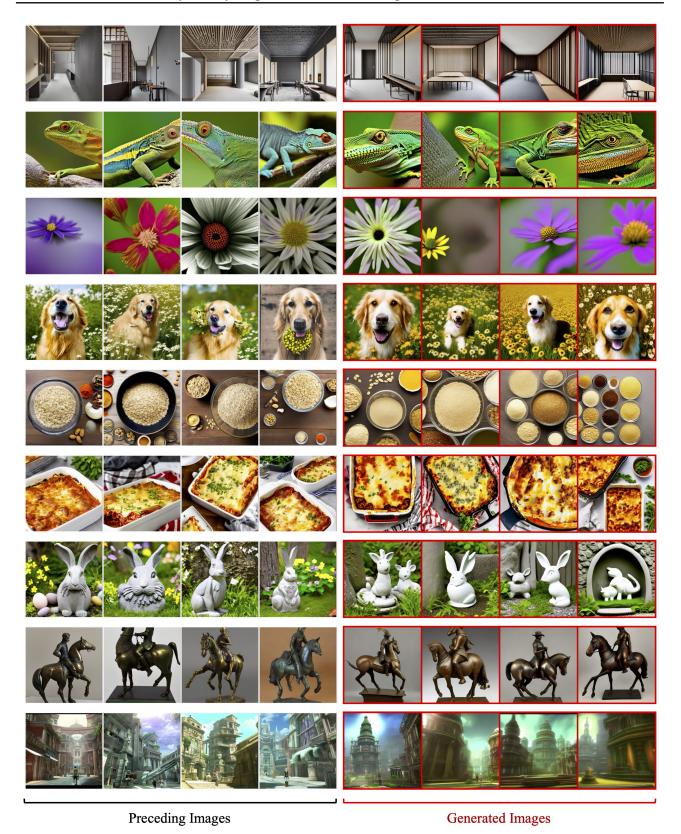
Preceding Images   Generated Images

Figure 10: **Images generated by M2M-Self.** Columns 1-4 showcase the preceding images for conditioning, and Columns 5-8 display the generated images.

Preceding Images

Generated Images

Figure 11: **(Continued) Images generated by M2M-Self.** Columns 1-4 showcase the preceding images for conditioning, and Columns 5-8 display the generated images.
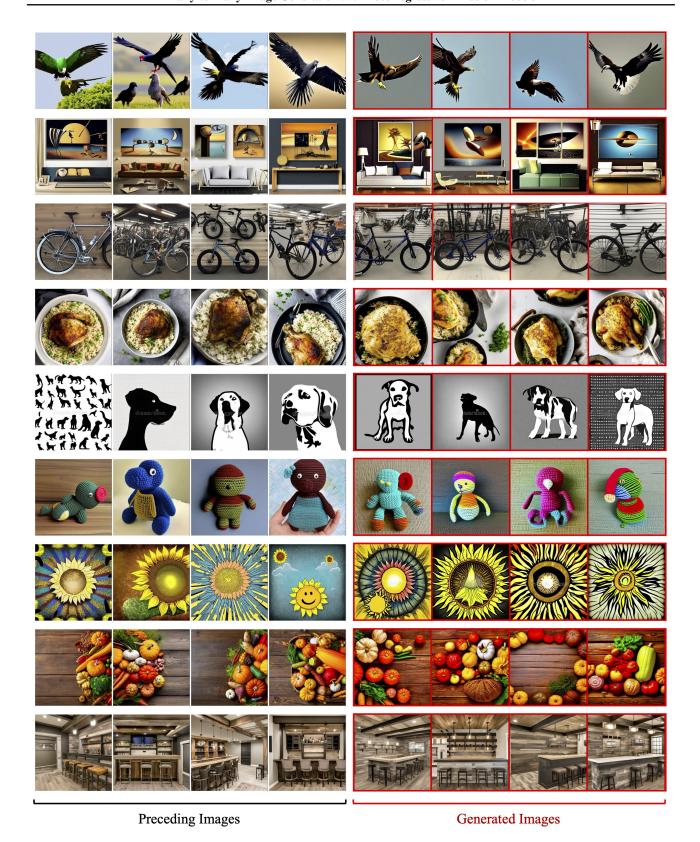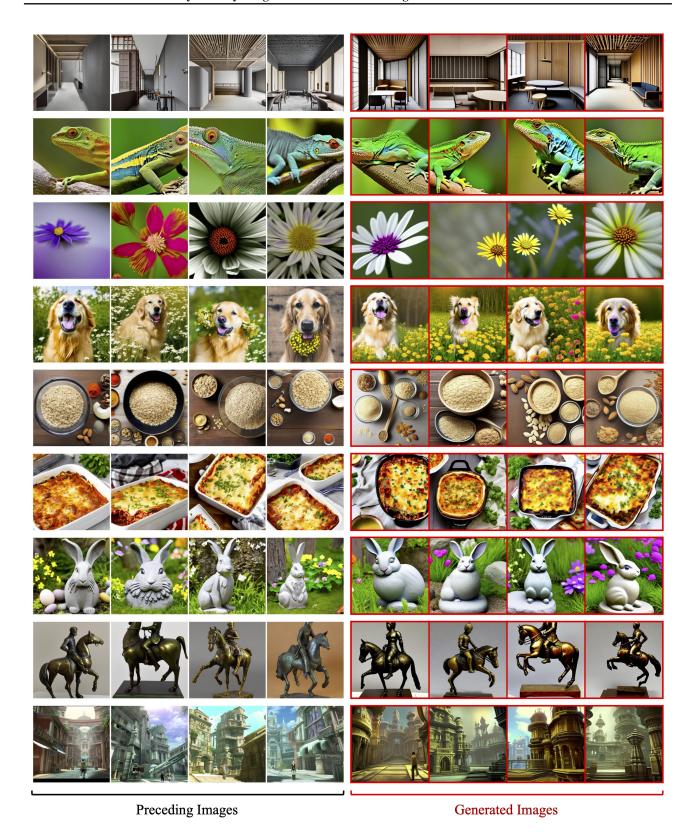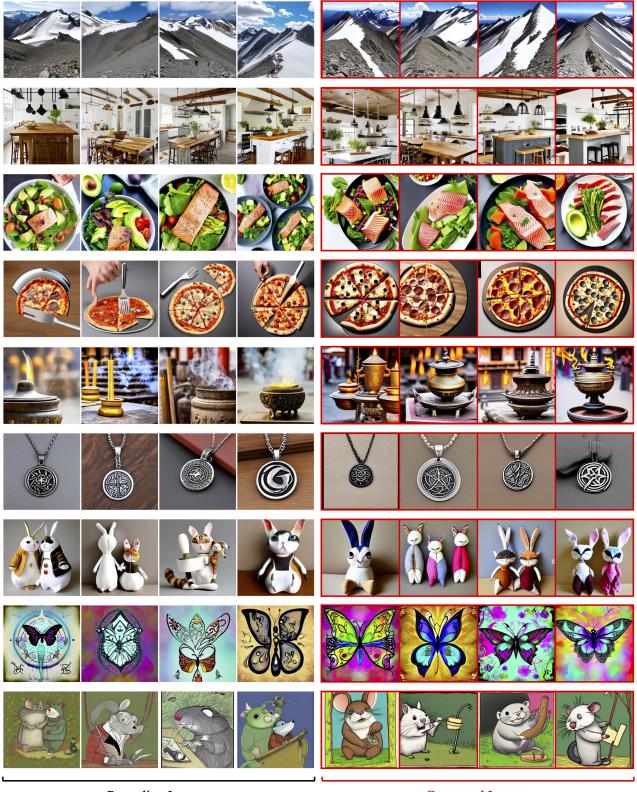
Preceding Images   Generated Images

Figure 12: **(Continued) Images generated by M2M-Self.** Columns 1-4 showcase the preceding images for conditioning, and Columns 5-8 display the generated images.

Figure 13: **Images generated by M2M-DINO.** Columns 1-4 showcase the preceding images for conditioning, and Columns 5-8 display the generated images.
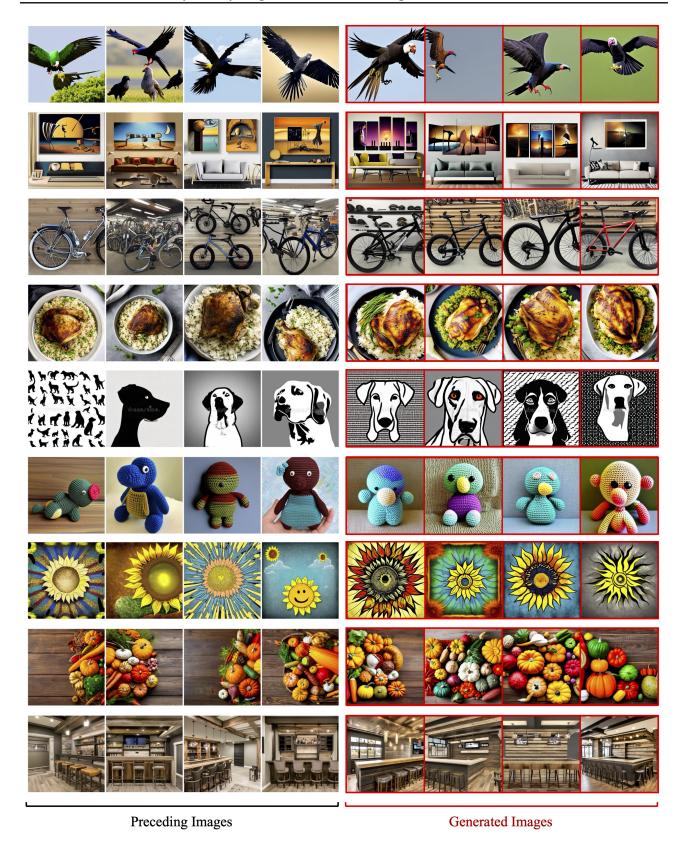
Preceding Images   Generated Images

Figure 14: **(Continued) Images generated by M2M-DINO.** Columns 1-4 showcase the preceding images for conditioning, and Columns 5-8 display the generated images.

Preceding Images   Generated Images

Figure 15: **(Continued) Images generated by M2M-DINO.** Columns 1-4 showcase the preceding images for conditioning, and Columns 5-8 display the generated images.
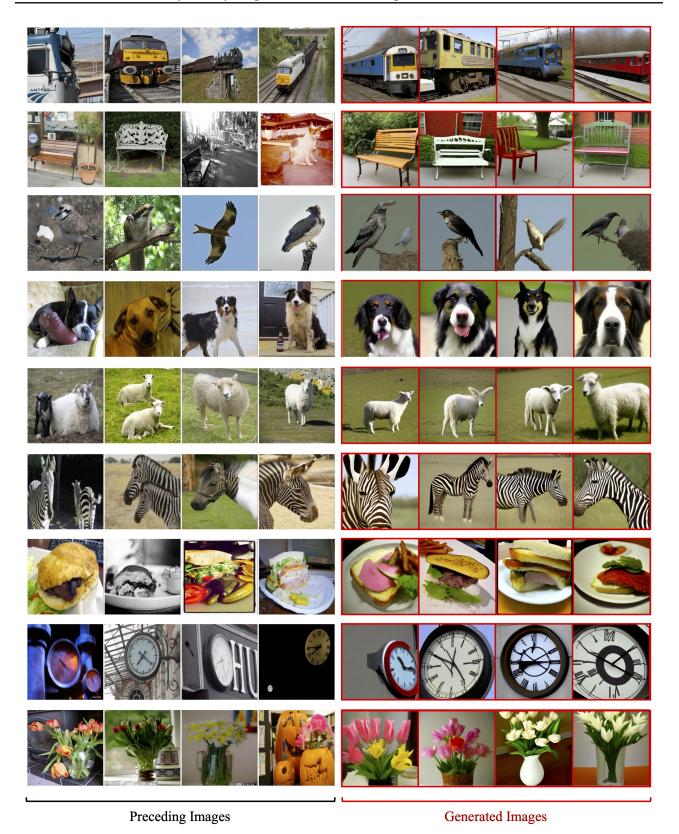
Figure 16: **Generalization to Real Images.** Columns 1-4 display the real images from the MSCOCO dataset, serving as the preceding images. Columns 5-8 showcase the corresponding images generated by **M2M-Self**, conditioned on the preceding images.
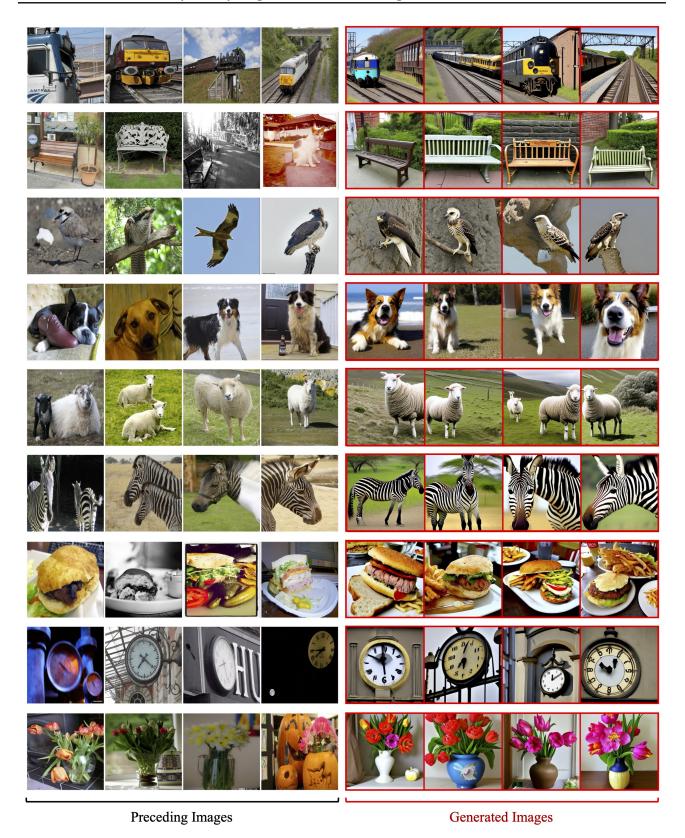
Figure 17: **Generalization to Real Images.** Columns 1-4 display the real images from the MSCOCO dataset, serving as the preceding images. Columns 5-8 showcase the corresponding images generated by **M2M-DINO**, conditioned on the preceding images.