Masked Generative Nested Transformers with Decode Time Scaling

Sahil Goyal [*1] **Debapriya Tula** [*†2] **Gagan Jain** [1] **Pradeep Shenoy** [1] **Prateek Jain** [1] **Sujoy Paul** [*1]

Abstract

Recent advances in visual generation have made significant strides in producing content of exceptional quality. However, most methods suffer from a fundamental problem - a bottleneck of inference computational efficiency. Most of these algorithms involve multiple passes over a transformer model to generate tokens or denoise inputs. However, the model size is kept consistent for all iterations, making it computationally expensive. In this work, we aim to address this issue primarily through two key ideas - (a) not all parts of the generation process need equal compute, hence we design a decode time model scaling schedule to utilize compute effectively, and (b) we can cache and reuse some of the intermediate computation. Combining these two ideas leads to using smaller models to process more tokens while large models process fewer tokens. These different-sized models do not increase the parameter size, as they share parameters. We rigorously experiment with ImageNet256×256 , ImageNet128×128, UCF101, and Kinetics600 to showcase the efficacy of the proposed method for image/video generation and frame prediction. Our experiments show that with almost 3× less compute than baseline, our model obtains competitive performance.

1. Introduction

The last decade has witnessed tremendous progress in image and video generation, under diverse paradigms - generative adversarial networks (Brock, 2018; Sauer et al., 2022), denoising processes such as diffusion models (Ho et al., 2020; 2022b; Dhariwal & Nichol, 2021; Rombach et al., 2022; Gu et al., 2022), image generation via vector quantized tokenization (Razavi et al., 2019; Esser et al., 2021; Ge et al., 2022; Van Den Oord et al., 2017), and so on. In recent years, diffusion models and modeling visual tokens as language have been the de-facto processes used to generate high-quality images. While initially proposed with a CNN or U-Net based architectures (Rombach et al., 2022; Saharia et al., 2022), transformer models have become the norm for these methods (Peebles & Xie, 2023; Yu et al., 2023a).

The recent advancements in visual generation can be categorized along two axes – (a) different types of denoising processes in the continuous latent space (Ho et al., 2020; Nichol & Dhariwal, 2021b), discrete space (Gu et al., 2022; Lou et al.) or masking in the discrete space (Yu et al., 2023a; Chang et al., 2022), continuous space (Li et al., 2024a) (b) modeling tokens either auto-regressively (Kondratyuk et al., 2024; Esser et al., 2021; Yu et al., 2021) with causal attention or parallel decoding with bi-directional attention (Gu et al., 2022; Yu et al., 2023a; Chang et al., 2022; Zheng et al., 2022). To achieve a high synthesis fidelity, both, denoising in diffusion models, and raster scan based auto-regressive token modeling require several iterations.

Recently, parallel decoding of discrete tokens have shown promise in generating high quality images with few iterations - MaskGIT (Chang et al., 2022), MAGVIT (Yu et al., 2023a), MUSE (Chang et al., 2023), MaskBIT (Weber et al., 2024), TiTok (Yu et al., 2024b). These models are trained with Masked Language Modeling (MLM) type losses, and the generation process involves unmasking a few confident tokens every decoding iteration, starting from all masked tokens. They can even surpass diffusion models, given a good visual tokenizer (Yu et al., 2023b; Weber et al., 2024).

Although MaskGIT reduces decode complexity significantly, parallel decoding still includes several redundant computations. First, the need for same capacity model for all steps needs to be investigated. Second, unlike auto-regressive models, which cache its computation in all steps, parallel decoding performs re-computation for all tokens. We empirically find that a smaller model can generate good-quality images quite fast, but its performance saturates after a point with more decoding iterations. A bigger model can perform finer refinement and generate better-quality images.

Motivated by these observations, we present **Ma**sked **G**enerate **Ne**sted **T**ransformers with Decode Time **S**caling (MaGNeTS). We design a model size curriculum over the

[*]Equal contribution , [†]work done while at Google DeepMind [1]Google DeepMind [2]University of California, Los Angeles. Correspondence to: Sahil Goyal <goyalsahil@google.com>, Sujoy Paul <sujoyp@google.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

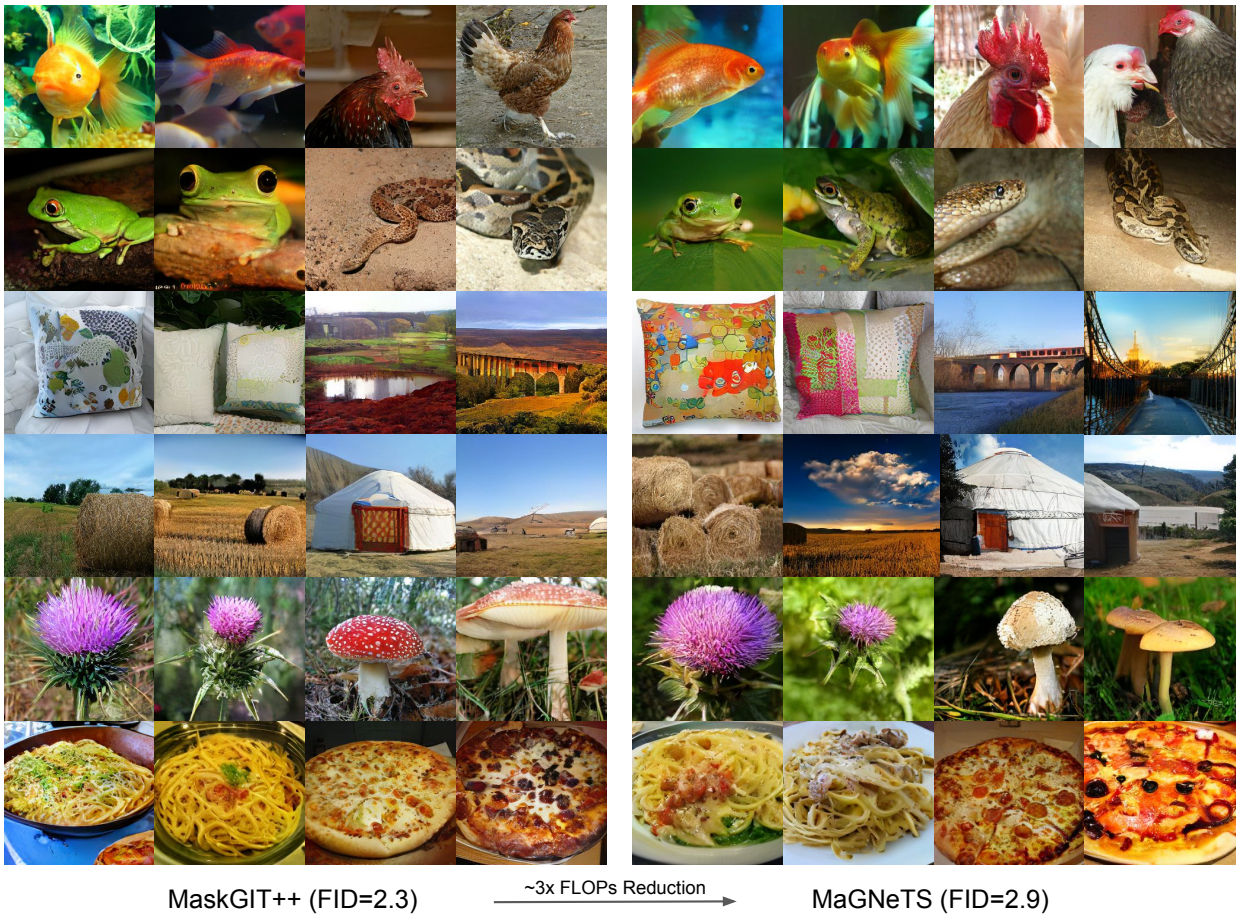


Figure 1: **Class-conditional image generation on ImageNet256×256**. Comparing MaskGIT++ and MaGNeTS (size: L).

decoding process, which efficiently utilize compute. MaGNeTS gradually scales the model size up to the full model over the decoding iterations instead of using a single large model throughout. Operating on discrete tokens, we cache key-value pairs of unmasked tokens and reuse them in later iterations. A combined effect of these two techniques leads to processing more tokens with smaller and fewer tokens with larger models. The heterogenous sized models share parameters as in MatFormer (Kudugunta et al., 2023). We build MaGNeTS on top of MaskGIT. We find that MaskGIT can be drastically improved using classifier-free guidance, specifically when trained with it. We call this MaskGIT++ and use this as the improved baseline, presenting all results on top of it. We also conduct preliminary inference-time experiments on diffusion models to demonstrate our method's generalization capabilities.

On ImageNet, with ~ 3× less compute, MaGNeTS generates images of similar quality as MaskGIT++ (see Figure 1). It is also comparable to state-of-the-art methods, which need orders of magnitude more compute. We also show MaGNeTS's efficacy on video datasets like UCF101 (Soomro et al., 2012) and Kinetics600 (Carreira et al., 2018). To summarize, the main contributions of this work are:

- We introduce the concept of model size scheduling during the generation process to significantly reduce compute requirements.
- We show that like auto-regressive models, KV-caching can also be used in parallel decoding, which can effectively reuse computation when refreshed appropriately.
- We introduce nested modeling in image/video generation to exploit the above ideas effectively.
- Extensive experiments show that MaGNeTS offers 2.5 - 3.7× compute gains across tasks.

2. Related Work

Efficient Visual Generation. Image generation literature has seen significant improvements in the past years - generative adversarial networks (Brock, 2018; Sauer et al., 2022), discrete token based models (Chang et al., 2022; Yu et al., 2023a), diffusion-based models (Kingma & Gao, 2023;

Hoogeboom et al., 2023), and more recently hybrid models (Peebles & Xie, 2023; Yu et al., 2024c), but they often guzzle computing power. Researchers tackle this bottleneck of computational costs with efficient model architectures and smarter sampling strategies.

In diffusion model literature, there have been some work to reduce the number of sampling steps, by treating the sampling process like ordinary differential equations (Song et al., 2022; Lu et al., 2022; Liu et al., 2022), incorporating additional training process (Kong & Ping, 2021; Nichol & Dhariwal, 2021a; Salimans & Ho, 2022; Song et al., 2023), sampling step distillation (Salimans & Ho, 2022; Song et al., 2023; Berthelot et al., 2023; Meng et al., 2023; Feng et al., 2024), sampling and training formulation modifications (Esser et al., 2024; Song et al., 2023), and more. Recently, there has been growing interest in understanding how each step in the diffusion sampling process contributes (Choi et al., 2022; Park et al., 2023; Lee et al., 2024). These approaches analyze sampling steps leveraging distance metrics such as LPIPS, fourier and spectral density analysis. Building on these explorations researchers have designed methods based on optimal sampling steps (Watson et al., 2022; Lee et al., 2024), weighted training loss (Choi et al., 2022), and step-specific models (Li et al., 2023; Yang et al., 2024; Lee et al., 2023). These step-specific models use computationally expensive evolutionary search algorithms, directly optimizing the quality metric, FID (Heusel et al., 2017). Concurrently, researchers are actively addressing the inherent architectural costs of diffusion models, particularly those associated with transformer attention mechanisms (Yuan et al., 2024; Yan et al., 2024).

On the other hand, certain works focus on building efficient and improved tokenizers. LDM (Rombach et al., 2022) takes diffusion models from pixel to compressed latent space for efficient and scalable generation. FSQ (Mentzer et al., 2023), LFQ (Yu et al., 2023b) and BSQ (Zhao et al., 2024) explore certain vector quantization techniques in the discrete tokenization process without explicitly learning the codebooks. VAR (Tian et al., 2024) explores multi-scale tokenizer to improve the generation quality. Recently TiTok (Yu et al., 2023b), FlowMo (Sargent et al., 2025) relax the topology of latent space and reduce the grid dimensionality of latents from 2D to 1D. Further FlexTok (Bachmann et al., 2025) enables adaptive sequence length in 1D tokenization for efficient generation. Instead of sampling or tokenization process optimization, we tackle an orthogonal problem of efficient compute allocation over the multi-step generation process. This makes our approach usable with a variety of tokenizers, model architectures and sampling schemes.

Nested Models. Rippel et al. (2014) introduced nested dropout to learn ordered representations of data that improve retrieval speed and adaptive data compression. Matryoshka Learning (Kusupati et al., 2022) introduces the concept of nested structures into embedding dimensions, making them elastic. MatFormer (Kudugunta et al., 2023) applies the same concept to the MLP hidden layer in each transformer block, enabling extraction of multiple sized models. MoNE (Jain et al., 2024) and Flextron (Cai et al., 2024) learn to route tokens to variable sized nested models leading to compute efficient processing. Stochastic Bottleneck (Koike-Akino & Wang, 2020), MQT (Hu et al., 2024), One-D-Piece (Miwa et al., 2025), FlexTok (Bachmann et al., 2025) and Semanticist (Wen et al., 2025) explore nesting across the latent sequence dimension. In this work, we demonstrate how different stages of a multi-step task like image/video generation, can be efficiently handled by nested models instead of relying on the full model at every step, significantly reducing computation without sacrificing quality.

3. Preliminaries

Parallel Decoding for Image Generation. Masked Generative Image Transformer (MaskGIT) (Chang et al., 2022) introduces a novel approach to image generation that significantly differs from traditional autoregressive models. In autoregressive decoding, images are generated sequentially, one pixel/token at a time, following a raster scan order (Esser et al., 2021; Kondratyuk et al., 2024; Wang et al., 2024; Yu et al., 2024a; Li et al., 2024b). This sequential approach is computationally inefficient, as each token is conditioned only on the previously generated tokens, leading to a bottleneck in processing time. MaskGIT generates all tokens of an image simultaneously, while iteratively refining them. This method enables significant acceleration in the decoding process. The tokens are discrete and obtained using Vector Quantized (VQ) autoencoders, learned with self-reconstruction and photo-realism losses (Yu et al., 2023a). The iterative parallel decoding process is represented as:

$$\mathbf{X}_k \leftarrow \text{Mask} \circ \text{Sample}(M(\mathbf{X}_{k-1}, c), k) \tag{1}$$

where $\mathbf{X} \in \mathbb{Z}_{\geq 0}^{N}$, are the input tokens, N is the number of tokens, $k \in [1, K]$ denote the iteration number, with K being the total number of iterations, $\mathbf{X}_0$ is either completely masked for full generation, and partially masked for conditional generation tasks like frame prediction, c is the category of image/video under generation. The Sample function utilizes logits predicted by the model $M(.)$, introduces certain randomness, and sorts them by confidence, unmasking only top-k tokens while masking the rest. We follow this process as in (Chang et al., 2022; Yu et al., 2023a).

Nested Models. The core of our algorithm for inference-efficient decoding relies on variable-sized nested models for efficient parameter-sharing. We use MatFormer's (Kudugunta et al., 2023) modeling approach to extract mul-

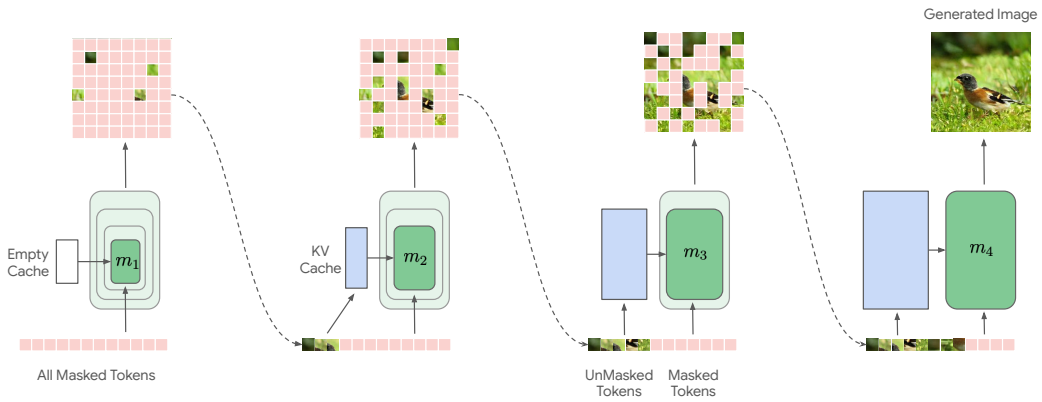


Figure 2: **MaGNeTS Decoding.** We start from the smallest nested model with an empty cache and gradually move to bigger models over the decoding iterations. We iterate using a particular sized model for a few iterations, before moving onto the next model size. As we cache the key-value pairs for the unmasked tokens, the KV cache size also increases over time. We also refresh the cache when we switch models, hence its dimension also increases over decoding iterations.

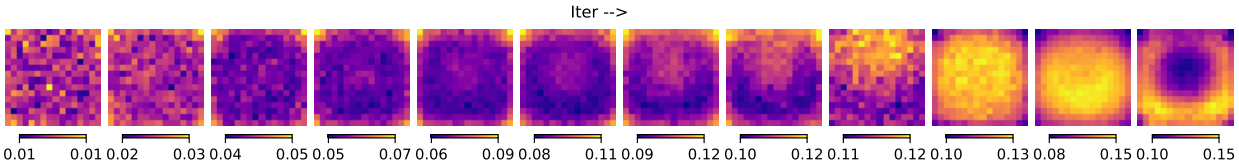


Figure 3: **Unmasked Token Density** visualization in each decoding iteration averaged over 50k generated samples on ImageNet. Yellow represents higher density. Each pixel represents a token from 16×16 latent token space. (See Appendix A for category-wise token density).

tiple nested models, from a single model, without increasing the total parameter count. Given a full transformer model M, MatFormer defines nested models $\{m_1, \ldots, m_C\}$, such that $m_1 \subset m_2 \cdots \subset m_C = M$. Each m_i has fewer parameters and reduced compute. The core idea of extracting nested models is that in a transformer block, a reduced computation using a parameter subspace is performed via a sliced matrix multiplication. Assuming a parameter matrix $\mathbf{W} \in \mathbb{R}^{d' \times d}$ and feature vector $\mathbf{x} \in \mathbb{R}^d$, then the computation $\mathbf{y} = \mathbf{W}\mathbf{x}$ is partially obtained by computing $\mathbf{y}_{[:\frac{d'}{p}]} = \mathbf{W}_{[:\frac{d'}{p},:]}\mathbf{x}$, if $\mathbf{y}$ is desired to be partial and $\mathbf{y} = \mathbf{W}_{[:,:\frac{d}{p}]}\mathbf{x}_{[:\frac{d}{p}]}$, if input $\mathbf{x}$ is partial. With such partial computations throughout the network, we can obtain nested models which share parameters.

While MatFormer (Kudugunta et al., 2023) obtained submodels with partial computation only in the MLP layer, we also do it in the Self-Attention layer, specifically in obtaining the $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ features. These features are of dimension $n_h \times \frac{d_h}{p}$, where n_h is the number of attention heads, d_h is the head feature dimension, and p is the model downscaling factor. We choose four downscaled models $C = 4$, with $p \in \{1, 2, 4, 8\}$ in this work. After attention computation, this gives us features that are also p times downscaled. Then, it is projected back to the full model dimension d using partial computation, as the input features are partial. The same strategy is applied to the MLP layer. This process gives us models with close to linear reduction in parameter count and inference compute with the downscaling factor p.

4. Method

Given the preliminaries, here we introduce the core algorithm. We first discuss the idea of scheduling models of different sizes over decode iters of MaskGIT. Then, we discuss the process of caching key-value in parallel decoding, followed by how to refresh them to improve performance. We finally discuss the nested model training method. A pictorial overview of our method is presented in Figure 2.

Decode Time Model Schedule. In iterative parallel decoding (Chang et al., 2022; Yu et al., 2023a), the same-sized model is used for all steps, starting with all tokens being masked. However, we hypothesize that certain stages of the generation process might be easier than others. For example, in the initial steps, the model only needs to capture coarse global structures, which can be achieved efficiently using smaller models. In the later steps, the model must refine finer details, which requires larger models. This hypothesis is bolstered with Figure 3, which shows that the generation process starts unmasking tokens from the background and shifts to the middle of the image in the later iterations (more

categorical examples in Appendix Figure 8).

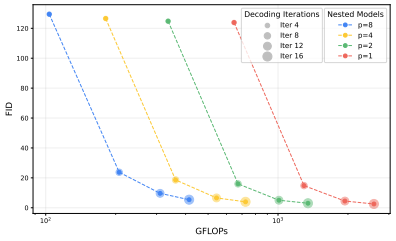


Figure 4: **Nested Models at different decoding iterations**. Different values of the downscaling factor p correspond to the nested models. The diameter of the blobs indicates #iterations.

Our hypothesis is further motivated by Figure 4, which presents the generation quality (FID) over iterations of parallel decoding for different-sized models. The smallest model reaches a reasonably good FID score with very low FLOPs compared to the biggest model. However, it saturates after a point, and the larger models surpasses the smaller ones in performance, demonstrating their ability to capture finer details and generate higher-quality images when provided with sufficient compute. This trend suggests that dynamically scaling the model size during decoding can exploit the varying task difficulty and achieve compute efficiency.

We use nested models to extract multiple models rather than using models with disjoint parameters. Nested models do not increase the parameter count and it also helps in better alignment of hypothesis when we shift model size over decode steps. The decode time model schedules can be generalized and represented as making the model choice in Equation (1) dependent on the iteration index as follows:

$$\mathbf{X}_k \leftarrow \text{Mask} \circ \text{Sample}(\mathcal{M}_k(\mathbf{X}_{k-1}, c), k)$$
$$\mathcal{M} = \{(m_{p_1})^{k_1}, (m_{p_2})^{k_2}, \ldots, (m_{p_n})^{k_n}\}, \quad \text{s.t.} \sum_i^n k_i = K \quad (2)$$

where $p_1, p_2, \ldots, p_n$ denote the downscaling factors of the corresponding nested models, and $(m)^k$ denotes that model m will be executed for k iterations. K represents the total number of iterations. We can think of different model schedules - (a) downscaling (starting with the full model and then gradually moving to the smallest model), (b) upscaling, (c) intermittently switching among a few models, and so on. We can also modify the integers k_i to choose the number of times we stick to a model before switching. However, as intuitively discussed before, we empirically validate that gradually upscaling the model size gets the best trade-off between the compute and generation quality.

Cached Parallel Decoding. Inspired by caching key-value pairs in auto-regressive models, we explore caching in parallel decoding, which retains relevant computations and enhances efficiency. In auto-regressive models, caching progressively happens in one fixed direction. However, in parallel decoding, caching must depend on which tokens are unmasked over the iterations.

Concretely, starting from an empty cached set, we keep adding keys and values to the set for the tokens that are unmasked after the Mask∘Sample steps (see Section 3). We do not update the predicted token indices for these unmasked tokens in the subsequent iterations. Hence, the cached key and values for the unmasked tokens are the only features the other masked tokens need. In every decoding iteration, we can categorize tokens into three main categories: unmasked tokens (for which we have cached KV), masked tokens that will be unmasked during the current iteration, and the rest of the masked tokens. Note that the KV cache for the second category tokens cannot be used in the next iteration but only in the iteration after that once we know their token indices in the current forward pass. We cache them in the next iteration for use in the immediately next iteration.

Caching is even more useful for decode time model schedules. For a schedule that progressively scales up the model size as decoding progresses, smaller models process more tokens, while the larger models process fewer tokens, leading to an efficient yet good quality image generation process.

Algorithm 1 MaGNeTS Decoding Algorithm

Input: $\mathbf{X}_0$ (Initial Tokens), K (#steps), N (#tokens), $\mathcal{M}$ (Nested Model Schedule), c (class),
Initialize: $k \leftarrow 0$; $cache \leftarrow \{\}$;
Note: $\mathbf{X}_0$ is a list of token ids (Mask token id = -1)
while *step* $k < K$ **do**
  if $k > 0$ *and* $\mathcal{M}_k \neq \mathcal{M}_{k-1}$ **then**
    Clear $cache$
  end
  Get uncached tokens: $\mathbf{X}_k^{uc} \leftarrow \{x_i \mid x_i \in \mathbf{X}_k;\ i \notin cache\}$
  Compute prediction probabilities and key-values: $p^k, (kv) \leftarrow \mathcal{M}_k(\mathbf{X}_k^{uc}; cache)$
  Sample tokens using current predictions p^k, without modifying previous predictions,
  $$\mathbf{X}^{k+1} \leftarrow \text{MaskGIT-Sample}(p^k)$$
  New indices to cache: $\mathcal{C} \leftarrow \{i \mid i \notin cache,\ \mathbf{X}_i^{k+1} \neq -1\}$
  Update the kv cache: $cache \leftarrow cache \cup \{i : (kv)_i \mid i \in \mathcal{C}\}$
  $k \leftarrow k + 1$
end
return $\mathbf{X}^K$

Intermittent Cache Refresh. Caching the key-value pairs for the unmasked tokens helps reduce computation, but it can slightly degrade performance. This happens because - (a) when we cache, the unmasked tokens are not updated in the subsequent iterations. (b) when we shift model size during generation, in the attention layer, the query size differs from the cached KV (see Section 3). While technically, we can zero-pad the KV to be compatible with the current model's query dimension, the model remains unfamiliar of such feature discrepancies between query and key-value.

To remedy this, we strategically refresh the cache while changing the model size. Refreshing involves discarding the cached KV for that iteration and caching a newly com-

puted KV for the immediate next iteration. We empirically find that it bridges the performance gap that arises due to caching. The proposed decode time model scaling algorithm is presented in Algorithm 1, which uses MaskGIT's sampling strategy (Chang et al., 2022; Yu et al., 2023a) to sample tokens from logits predicted by the network.

Training Nested Models. MatFormer (Kudugunta et al., 2023) opts for a joint optimization of losses w.r.t. ground-truth from all models with equal weights. While this mode of training works for a small range of model downscaling, we found it to hurt performance with larger downscaling factors p. We introduce a combination of ground truth and distillation loss to address this issue. We perform online distillation progressively, where the teacher for model m_i is model m_{i+1}. The full model $m_N (= M)$ is trained with only ground truth loss. This provides a simpler optimization for the smaller nested models while maintaining the overall objective. Progressive distillation also reduces the teacher-student size gap, which can otherwise hurt distillation performance (Stanton et al., 2021; Beyer et al., 2022; Mirzadeh et al., 2019). Given input $\mathbf{X}$, ground truth label $\mathbf{Y}$ and loss function $\mathcal{L}$, our training loss is expressed as:

$$\mathcal{L}_{train} = \frac{1}{N}\Big(\mathcal{L}(m_N(\mathbf{X}), \mathbf{Y}) + \sum_{i=1}^{N-1} \alpha_i \mathcal{L}(m_i(\mathbf{X}), \mathbf{Y}) + (1-\alpha_i)\mathcal{L}(m_i(\mathbf{X}), m_{i+1}(\mathbf{X}))\Big) \tag{3}$$

where α_i controls the weight between the distillation and ground truth loss, which is linearly decayed from 1 to 0 as training progresses. Note that a stop gradient is applied during distillation on m_{i+1} in the third term of the equation.

Classifier-Free Guidance. Following literature (Ho & Salimans, 2022; Yu et al., 2023b), we also utilize classifier-free guidance during the generation process. Following the same motivation as decode time model scaling discussed above, which shows that the initial decoding iterations focus on the background region, and gradually moves to the main object/region of interest in the final decoding iterations, we apply guidance to only a few final decoding iterations. We find that doing this offers similar quality images as applying guidance to all iters (refer Figure 9b). See Appendix B for detailed analysis.

5. Experiments and Results

We conduct extensive experiments to demonstrate the efficacy of our approach on three distinct tasks: class-conditional image generation, class-conditional video generation, and frame prediction.

Datasets. We evaluate our model on ImageNet256×256 and ImageNet128×128 (Deng et al., 2009) for image generation, UCF101 (Soomro et al., 2012) for video generation and Kinetics600 (Carreira et al., 2018) for frame prediction (5-frame condition).

Implementation Details. We utilize the pretrained tokenizers from MaskGIT (Chang et al., 2022) (for images) and MAGVIT (Yu et al., 2023a) (for videos) with the codebook size of 1024 tokens. We train different models for image sizes 256×256 and 128×128. Respective tokenizers compress them to 16×16 discrete tokens. For videos, we learn models for $16 \times 128 \times 128$, where the tokenizer outputs $4 \times 16 \times 16$ tokens. Following MaskGIT, we utilize the Bert model (Devlin et al., 2019) as a transformer backbone. We perform experiments at several model scales to understand the scaling behaviors of our algorithm. We utilize the same training hyper-parameters to train our nested models as these baselines. We train our model for 270 epochs for all the experiments. Unless otherwise mentioned, throughout the paper, we employ same number of steps per model before switching to the next model, i.e., $k_1 = k_2 = ... = k_n$. We follow a cosine schedule of unmasking tokens during inference. For image generation and frame prediction, we use classifier-free guidance for both MaGNeTS and respective baselines. We drop input class condition labels for 10% of the training batches in image generation to better facilitate classifier-free guidance during image generation. We mention the details of sampling hyperparameters in Appendix B.

Evaluation Metrics. Following previous baselines, we use Fréchet Inception Distance (FID) (Heusel et al., 2017; Dhariwal & Nichol, 2021) for image generation, Fréchet Video Distance (FVD) (Unterthiner et al., 2019) for the video generation tasks, Inception Score (Salimans et al., 2016) for both tasks, as well as precision and recall for image generation. We compare algorithms using inference-time GFLOPs. Refer Appendix D for GFLOPs computation details.

5.1. Image Generation

Comparison with Baselines. In this section, we compare MaGNeTS with state-of-the-art methods in the literature for image generation. We list the results for 256×256 and 128×128 image generation on ImageNet-1k in Table 1 and Table 2 respectively. Table 1 shows that MaGNeTS can speed up the generation process by $2.65 - 3\times$ (depending on total step count) compard to MaskGIT++, with a negligible drop in FID. Refer Appendix D for real-time gains. Figure 5 illustrates that MaGNeTS significantly accelerates parallel decoding, which gets more pronounced as image resolution grows. Figure 1 and Figure 10 show generated images from MaskGIT++ and MaGNeTS (ours). As shown in recent literature, using a superior tokenizer (Yu et al., 2023b; Weber et al., 2024) or optimized training/inference configu-

Model	AR	FID ↓	IS ↑	Prec ↑	Rec ↑	# params	# steps	# Gflops
BigGAN-deep$^{\square}$ (Brock, 2018)		7.0	171.4	87	28	160M	1	-
StyleGAN-XL$^{\square g}$ (Sauer et al., 2022)		2.3	265.1	-	-	166M	1	-
Improved DDPM$^{\square}$ (Nichol & Dhariwal, 2021b)		12.3	-	70	62	280M	250	>150k
ADM + Upsample$^{\square g}$ (Dhariwal & Nichol, 2021)		3.9	215.8	83	53	554M	250	371k
LDM-4$^{\square g*}$ (Rombach et al., 2022)		3.6	247.7	-	-	400M	250	51.5k
DiT-XL/2$^{\square g*}$ (Peebles & Xie, 2023)		2.3	278.2	83	57	675M	250	59.5k
MDT$^{\square g*}$ (Gao et al., 2023)		1.8	283.0	81	61	676M	250	>59k
MaskDiT$^{\square g*}$ (Zheng et al., 2023)		2.3	276.6	80	61	736M	250	>28k
CDM$^{\square}$ (Ho et al., 2022a)		4.9	158.7	-	-	-	8100	-
RIN$^{\square}$ (Jabri et al., 2022)		3.4	182.0	-	-	410M	1000	334k
Simple Diffusion$^{\square g}$ (Hoogeboom et al., 2023)		2.4	256.3	-	-	2B	512	-
VDM++$^{\square g}$ (Kingma & Gao, 2023)		2.1	267.7	-	-	2B	512	-
EDiff$^{\square g}$ (Hang et al., 2024)		2.1	-	-	-	450M	50	119k
LPDM-ADM$^{\square g}$ (Wang et al., 2023)		2.7	-	-	-	-	50	7.8k
MAR$^{\square g}$ (Li et al., 2024b)	✓	1.8	296.0	81	60	479M	128	-
VQVAE-2$^{\square}$ (Razavi et al., 2019)	✓	31.1	~45	36	57	13.5B	5120	-
VQGAN$^{\square}$ (Esser et al., 2021)	✓	15.8	78.3	-	-	1.4B	256	-
VQGAN (architecture) + MaskGIT (setup)$^{\square}$		18.7	80.4	78	26	227M	256	-
MaskGIT$^{\square}$(Chang et al., 2022)		6.2	182.1	80	51	227M	8	647
Mo-VQGAN$^{\square}$ (Zheng et al., 2022)		7.2	130.1	72	55	389M	12	~1k
MaskBit$^{\square g}$ (Weber et al., 2024)		1.7	341.8	-	-	305M	64	10.3k
PAR-4×$^{\square}$ (Wang et al., 2024)	✓	3.8	218.9	84	50	343M	147	-
PAR-16×$^{\square}$ (Wang et al., 2024)	✓	2.9	262.5	82	56	3.1B	51	-
MaskGIT++^{g_4}		2.5	260.3	83	54	303M	12	1.3k
MaskGIT++^{g_6}		2.3	280.6	84	51	303M	16	1.8k
MaGNeTS (ours)^{g_4}		3.1	254.8	85	50	303M	12	490
MaGNeTS (ours)^{g_6}		2.9	253.1	84	51	303M	16	608

Table 1: **Class-conditional Image Generation** on ImageNet 256×256. "# steps" refers to the number of neural network runs. $^{\square}$ denotes values taken from prior publications. * indicates usage of extra training data. g denotes use of classifier-free guidance (Ho & Salimans, 2022) for all steps. g_x represents use of guidance only for final x steps.

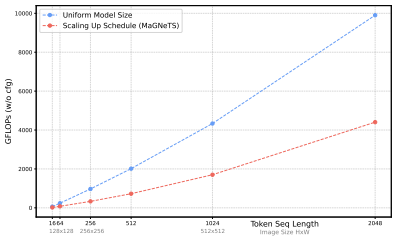


Figure 5: **Compute Comparison** between uniform model schedule (MaskGIT) and MaGNeTS, for 12 decode iters.

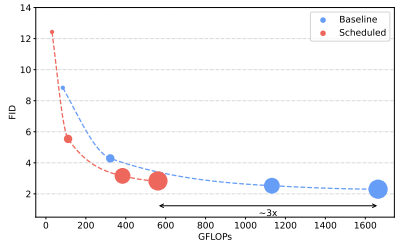


Figure 6: **Compute Scaling Curve.** Generation performance vs compute for different model sizes. The blob size indicates parameter count.

rations (Ni et al., 2024a;b) can further boost MaGNeTS's performance.

Note that, a direct comparison with several recent diffusion methods isn't feasible, as they typically report results on ImageNet64×64. We report some numbers in Table 1 for diffusion models at 256×256 resolution for comparison, though they may use different tokenizers or generation architectures. Exploring these variations is beyond the scope of this work. In addition to our experiments with non-autoregressive transformers, we conduct preliminary experimentation on applying the core idea of decode-time model scaling to diffusion transformers (see Appendix C.1).

Method	FID	# params	# steps	# GFLOPs
DPM-Solver$^{\square g}$ (Lu et al., 2022)	4.1	422M	12	>3k
MaskGIT++^{g_4}	3.2	303M	12	1.3k
MaGNeTS (ours)^{g_4}	3.9	303M	12	490

Table 2: **Class-conditional Image Generation** on ImageNet128×128. "# steps" refers to the number of neural network runs. $^{\square}$ denotes values taken from prior publications. g denotes use of classifier-free guidance (Ho & Salimans, 2022) for all steps. g_x represents use of guidance only for final x steps.

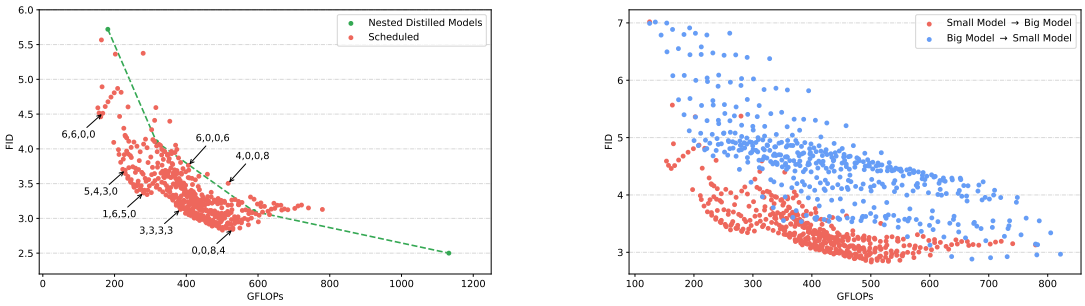
Scaling Analysis. To understand the scaling properties of MaGNeTS we train models of different sizes - S (22M), B (86M), L (303M) and XL (450M) for both the baseline as well as nested models needed for our algorithm. We use the same hyper-parameters for all, such as learning rate, epochs, weight decay, etc. We present the results in Figure 6. It shows the compute vs performance of different models, with the blob size denoting the model size. For a certain parameter count, the baseline uses the full model for all 12 decoding steps, while the scheduled routines use a sequence of nested models with downsampling factors $p = 8, 4, 2, 1$ for 3 steps each. Scaling up model size lead to much cheaper compute scaling of MaGNeTS than the baseline, with almost 3× compute reduction.

5.2. Video Generation

We use the MAGVIT (Yu et al., 2023a) framework to train parallel decoding based video generation and frame prediction models. Figure 11 shows generated videos of UCF101. We summarize the results for class-conditional video generation on UCF101 in Table 3 and for frame prediction on Kinetics600 in Table 4. Despite the challenging nature of video generation relative to image generation, results indicate that the decode time scaling of model size holds true even for video generation. MaGNeTS remains competitive to MAGVIT for frame prediction with ~ 3.7× lower compute.

5.3. Ablation Studies

Impact of Decode Time Model Schedule. We study the effect of different model scheduling choices. As discussed previously, we can think of different model schedules - scaling up model size, scaling down, periodic scaling up and



(a) Scaling Up Schedules

(b) Scaling up vs down

Figure 7: **Scheduling Options.** (a) This shows the compute-performance trade-off for different schedule options while always scaling up model size over generation iters. The four numbers for each point denote the number of iters each model size operates in the order of downsampling factor $p = (8, 4, 2, 1)$. (b) This shows the benefit of scaling up model size compared to scaling it down during decoding.

Method	Class	FVD ↓	IS↑	# params	# steps	# GFlops
RaMViD□* (Höppe et al., 2022)		-	21.71 ± 0.21	308M	500	-
StyleGAN-V□* (Skorokhodov et al., 2022)		-	23.94 ± 0.73	-	1	-
DIGAN□ (Yu et al., 2022)		577±21	32.70± 0.35	-	1	~148
DVD-GAN□ (Clark et al., 2019)	✓	-	32.97± 1.70	-	1	-
Video Diffusion□* (Ho et al., 2022b)			57.00± 0.62	1.1B	256	-
TATS□ (Ge et al., 2022)		420± 18	57.63± 0.24	321M	1024	-
CCVS+StyleGAN□ (Le Moing et al., 2021)		386± 15	24.47± 0.13	-	-	-
Make-A-Video□* (Singer et al., 2022)	✓	367	33.00	-	-	-
TATS□ (Ge et al., 2022)	✓	332± 18	79.28± 0.38	321M	1024	-
CogVideo□* (Hong et al., 2022)	✓	626	50.46	9.4B	-	-
Make-A-Video□* (Singer et al., 2022)	✓	81	82.55	≫3.5B	≫250	-
PAR-4×□ (Wang et al., 2024)	✓	99.5	-	792M	323	-
PAR-16×□ (Wang et al., 2024)	✓	103.4	-	792M	95	-
MAGVIT-B□ (Yu et al., 2023a)	✓	159± 2	83.55± 0.14	87M	12	~1.3k
MAGVIT-L (Yu et al., 2023a)	✓	74.4± 2	89.54± 0.21	306M	12	~4.3k
MaGNeTS (ours)	✓	96.4±2	88.53±0.20	306M	12	~1.7k

Table 3: **Class-conditional Video Generation** on UCF-101. Methods in gray are pretrained on additional large video data. Methods with ✓ in the Class column are class-conditional, while the others are unconditional. Methods marked with * use custom resolutions, while the others are at 128×128. □ denotes values taken from prior publications. No guidance is used for UCF101.

Method	FVD ↓	IS ↑	# params	# steps	# GFlops
CogVideo□ (Hong et al., 2022)	109.2	-	9.4B	-	-
CCVS□ (Le Moing et al., 2021)	55.0±1.0	-	-	-	-
Phenaki□ (Villegas et al., 2022)	36.4 ± 0.2	-	1.8B	48	-
TrIVD-GAN-FP□ (Luc et al., 2020)	25.7 ± 0.7	12.54 ± 0.06	-	1	-
Transframer□ (Nash et al., 2022)	25.4	-	662M	-	-
RaMViD□ (Höppe et al., 2022)	16.5	-	308M	500	-
Video Diffusion□ (Ho et al., 2022b)	16.2 ± 0.3	15.64	1.1B	128	-
MAGVIT-B□	24.5± 0.9	-	87M	12	~1.3k
MAGVIT-L	7.2 ± 0.1	16.48 ± 0.01	306M	12	~ 4.3k
MAGVIT-L^{g_2}	6.6 ± 0.1	16.29 ± 0.01	306M	12	~ 5.1k
MaGNeTS (ours)	10.8 ± 0.1	16.25 ± 0.02	306M	12	~1.2k
MaGNeTS (ours)^{g_2}	9.6 ± 0.1	16.25 ± 0.01	306M	12	~1.4k

Table 4: **Frame prediction** on K600. □ denotes values taken from papers. g_x denotes use of guidance only for final x steps.

down, and so on. For this analysis, we consider the L-sized model, with three nested models within it with parameter reduction by roughly $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$. We can denote the number of times these four models are called during decoding as (k_1, k_2, k_3, k_4), s.t., $\sum_{i=1}^{4} k_i = 12$. We drop the model notation of m_p in Equation (2) for simplicity and explicitly mention the model names in the text, as discussed next.

First, we evaluate all combinations of k_i for which we always scale up in Figure 7a in red and scale down in Figure 7b in blue. The green curve shows the performance of the individual nested models. We have the following observations - (1) for a certain compute budget, the scheduling of models over generation iterations (red dots) can offer better performance than using a single nested model (green curve) for all steps. (2) Models that have smoother transitions in nested models, such as $(3, 3, 3, 3)$ or $(0, 0, 8, 4)$, offer much better performance than the ones which has abrupt model transition such as $(6, 0, 0, 6)$ or $(3, 0, 0, 9)$, i.e., directly jumping from the smallest to the biggest model. (3) Figure 7b shows that scaling up nested model size offers much better performance than scaling down model size. This shows that bigger models are better utilized in the later iters.

Impact of Caching and Refresh. We now discuss the impact of caching and its refresh. For this analysis, we use a uniform model schedule: $k_1 = k_2 = k_3 = k_4 = 3$. We also perform caching and refresh on the baseline model, which has not been trained with any nesting with the same model applied for all iterations. For the baseline, we also refresh the cache at exactly the same steps as the scheduled model. We present the results in Table 5. The columns "Baseline" and "Scheduled" do not involve any cache. While caching degrades the performance a bit, refreshing it intermittently can avoid the degradation. While refresh does have some compute overhead, it does help significantly to bridge the quality gap. Scheduling of models with caching and refresh has the best compute-performance trade-off.

Algorithm	Baseline	+ Cache	+ Refresh	Scheduled	+ Cache	+ Refresh
FID	2.5	3.4	2.6	3.1	4.8	3.1
FLOP Gains (times)	1.0	1.3	1.2	2.1	3.5	3.0

Table 5: **Caching Ablation.** Only caching performs inferior, which is bridged by refreshing it. A scaling up model schedule with caching and its refresh offers the best compute-performance trade-off. Results are on ImageNet256×256 with L-sized model.

The efficiency of using nested models. In MaGNeTS we use nested models instead of separately trained smaller sized models. This has two advantages - (a) parameter sharing, which limits the number of parameters to just that of the full model, compared to $1.875\times(= 1+1/2+1/4+1/8)$ for disjoint models. Increasing the parameter count will increase memory requirements. (b) Nested models are trained efficiently in just a single training run. When trained with distillation, they generate better models than training standalone models (refer Table 8) of the same size as the nested models. For performance comparison, we trained standalone (L-sized) models of the same size as the nested models for both UCF101 and ImageNet. The results are presented in Table 6. Nested models can efficiently share parameters without loss in performance (ImageNet) and offer constraints that help in better performance (UCF101) than using standalone models.

Dataset	Nested Models	Standalone Models
ImageNet (FID)	3.1	3.1
UCF101 (FVD)	96.4	115.0

Table 6: **Nested vs Standalone Models.** This table presents the performance comparison between using nested models vs. standalone, independent models without parameter sharing in the decode-time scheduling algorithm of MaGNeTS.

Impact of number of nested models. We train different settings $p = \{1, 2\}$ (two models), $p = \{1, 2, 4\}$ (three models), $p = \{1, 2, 4, 8\}$ (four models), $p = \{1, 2, 4, 8, 16\}$ (five models), and $p = \{1, 2, 4, 8, 16, 32\}$ (six models). We observe that for all of these cases, the biggest model performance remains almost same. However, the performance of the smaller models degrades as shown in Table 7.

We hypothesize that the drop in performance of smaller models is due to their lower representational power. As we add more nested models, the task complexity of the shared representation increases, and burdens the mid-sized model. However, this drop in performance does not significantly impact the performance of model scheduling, as the larger models dominate the final results. Note that all of these results are on top of models trained with progressive distillation (see Section 4), which helps to retain the performance to some extent as shown in Table 8.

Impact of Distillation. We use two types of losses to train the nested models - loss w.r.t the ground-truth tokens and distillation loss using the next bigger model as the teacher. The weight between the two losses is also linearly interpolated from the former to the latter. We compare this training strategy with the two extremes – only ground truth loss and only distillation loss and present the results in Table 8. As we can see, using only distillation loss results in divergence. Using ground-truth loss is also inferior to linearly annealing.

Nested Attention Heads We also investigate nesting along the number of attention heads (n_h), applying the same partial computation strategy as discussed in Section 3. However, this performed worse than nesting along the head feature dimension (which we use for attention parameter nesting).

Inference Model Schedules	# Trained Nested Models →				
	2	3	4	5	6
$(0, 0, 0, 0, 0, 12)$	**2.3**	2.4	2.4	2.4	2.4
$(0, 0, 0, 0, 12, 0)$	**2.6**	2.7	2.7	2.8	2.8
$(0, 0, 0, 12, 0, 0)$	-	**3.4**	3.5	3.7	3.8
$(0, 0, 12, 0, 0, 0)$	-	-	**5.2**	5.7	6.3
$(0, 12, 0, 0, 0, 0)$	-	-	-	**8.9**	10.8
$(0, 0, 0, 0, 6, 6)$	**2.6**	2.6	2.6	2.7	2.7
$(0, 0, 0, 4, 4, 4)$	-	**2.7**	2.8	2.8	2.8
$(0, 0, 3, 3, 3, 3)$	-	-	**3.1**	3.2	3.2
$(0, 3, 3, 2, 2, 2)$	-	-	-	**6.3**	6.6

Table 7: **Ablation of number of nested models.** Experiments are on ImageNet256×256 on L-sized model.

Dataset	Training Algo.	$p = 1$	$p = 2$	$p = 4$	$p = 8$	Scheduled
ImageNet	Only GT	2.4	2.9	3.9	5.7	3.1
	Only Distill	⟵ Training Diverged ⟶				
	GT → Distill	2.4	**2.7**	**3.5**	**5.2**	3.1
UCF101	Only GT	80.0	101.3	143.8	221.8	112.6
	Only Distill	⟵ Training Diverged ⟶				
	GT → Distill	**78.3**	**91.2**	**115.4**	**164.4**	**96.4**

Table 8: **Distillation Ablation.** This shows the impact of different training losses used for the nested models on ImageNet256×256 (size: L) and UCF101 (size: L). Using only distillation diverges while using only ground-truth losses performs worse than our approach (third row), where we combine ground-truth and distillation losses with a linear decay from the former to the latter.

6. Conclusion

In this paper, we propose MaGNeTS, to allocate variable compute along image/video generations steps. We show that instead of always using the same sized model for all decoding steps, we can start from a model which is nested and fraction of its full size, and then gradually increase model size. This along with key-value caching in the parallel decoding paradigm obtains significant compute gains. We believe that our exploration of dynamic compute opens exciting new research directions in efficient generative models. In future works, we plan to explore token-dependent model schedules for further compute gains.

Impact Statement

Our work mainly focuses on improving the inference time compute efficiency of state-of-the-art visual generative models in literature. As the work have been conducted on publicly available datasets, we do not see any potential ethical or societal concerns.

References

Bachmann, R., Allardice, J., Mizrahi, D., Fini, E., Kar, O. F., Amirloo, E., El-Nouby, A., Zamir, A., and Dehghan, A. Flextok: Resampling images into 1d token sequences of flexible length, 2025. URL https://arxiv.org/abs/2502.13967.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A vit backbone for diffusion models, 2023. URL https://arxiv.org/abs/2209.12152.

Berthelot, D., Autef, A., Lin, J., Yap, D. A., Zhai, S., Hu, S., Zheng, D., Talbott, W., and Gu, E. Tract: Denoising diffusion models with transitive closure time-distillation, 2023. URL https://arxiv.org/abs/2303.04248.

Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., and Kolesnikov, A. Knowledge distillation: A good teacher is patient and consistent, 2022. URL https://arxiv.org/abs/2106.05237.

Brock, A. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Cai, R., Muralidharan, S., Heinrich, G., Yin, H., Wang, Z., Kautz, J., and Molchanov, P. Flextron: Many-in-one flexible large language model. *arXiv preprint arXiv:2406.10260*, 2024.

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., and Zisserman, A. A short note about kinetics-600, 2018. URL https://arxiv.org/abs/1808.01340.

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. Perception prioritized training of diffusion models, 2022. URL https://arxiv.org/abs/2204.00227.

Clark, A., Donahue, J., and Simonyan, K. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., and Rombach, R. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Feng, W., Yang, C., An, Z., Huang, L., Diao, B., Wang, F., and Xu, Y. Relational diffusion distillation for efficient image generation, 2024. URL https://arxiv.org/abs/2410.07679.

Gao, S., Zhou, P., Cheng, M.-M., and Yan, S. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23164–23173, 2023.

Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.-B., and Parikh, D. Long video generation with time-agnostic vqgan and time-sensitive transformer, 2022. URL https://arxiv.org/abs/2204.03638.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. Vector quantized diffusion model for text-to-image synthesis, 2022. URL https://arxiv.org/abs/2111.14822.

Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., and Guo, B. Efficient diffusion training via min-snr weighting strategy, 2024. URL https://arxiv.org/abs/2303.09556.

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, 2022. URL https://arxiv.org/abs/2207.12598.

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

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022a.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Hoogeboom, E., Heek, J., and Salimans, T. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., and Dittadi, A. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022.

Hu, W., Dou, Z.-Y., Li, L. H., Kamath, A., Peng, N., and Chang, K.-W. Matryoshka query transformer for large vision-language models, 2024. URL https://arxiv.org/abs/2405.19315.

Jabri, A., Fleet, D., and Chen, T. Scalable adaptive computation for iterative generation. *arXiv preprint arXiv:2212.11972*, 2022.

Jain, G., Hegde, N., Kusupati, A., Nagrani, A., Buch, S., Jain, P., Arnab, A., and Paul, S. Mixture of nested experts: Adaptive processing of visual tokens, 2024. URL https://arxiv.org/abs/2407.19985.

Kingma, D. P. and Gao, R. Understanding the diffusion objective as a weighted integral of elbos. *arXiv preprint arXiv:2303.00848*, 2, 2023.

Koike-Akino, T. and Wang, Y. Stochastic bottleneck: Rateless auto-encoder for flexible dimensionality reduction, 2020. URL https://arxiv.org/abs/2005.02870.

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., et al. Videopoet: A large language model for zero-shot video generation. *ICML*, 2024.

Kong, Z. and Ping, W. On fast sampling of diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2106.00132.

Kudugunta, S., Kusupati, A., Dettmers, T., Chen, K., Dhillon, I., Tsvetkov, Y., Hajishirzi, H., Kakade, S., Farhadi, A., Jain, P., et al. Matformer: Nested transformer for elastic inference. *arXiv preprint arXiv:2310.07707*, 2023.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Le Moing, G., Ponce, J., and Schmid, C. Ccvs: Context-aware controllable video synthesis. *Advances in Neural Information Processing Systems*, 34:14042–14055, 2021.

Lee, H., Lee, H., Gye, S., and Kim, J. Beta sampling is all you need: Efficient image generation strategy for diffusion models using stepwise spectral analysis, 2024. URL https://arxiv.org/abs/2407.12173.

Lee, Y., Kim, J.-Y., Go, H., Jeong, M., Oh, S., and Choi, S. Multi-architecture multi-expert diffusion models, 2023. URL https://arxiv.org/abs/2306.04990.

Li, L., Li, H., Zheng, X., Wu, J., Xiao, X., Wang, R., Zheng, M., Pan, X., Chao, F., and Ji, R. Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration, 2023. URL https://arxiv.org/abs/2309.10438.

Li, T., Tian, Y., Li, H., Deng, M., and He, K. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024a.

Li, T., Tian, Y., Li, H., Deng, M., and He, K. Autoregressive image generation without vector quantization, 2024b. URL https://arxiv.org/abs/2406.11838.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds, 2022. URL https://arxiv.org/abs/2202.09778.

Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, 2022. URL https://arxiv.org/abs/2206.00927.

Luc, P., Clark, A., Dieleman, S., Casas, D. d. L., Doron, Y., Cassirer, A., and Simonyan, K. Transformation-based adversarial video prediction on large-scale data. *arXiv preprint arXiv:2003.04035*, 2020.

Meng, C., Rombach, R., Gao, R., Kingma, D. P., Ermon, S., Ho, J., and Salimans, T. On distillation of guided

diffusion models, 2023. URL https://arxiv.org/abs/2210.03142.

Mentzer, F., Minnen, D., Agustsson, E., and Tschannen, M. Finite scalar quantization: Vq-vae made simple, 2023. URL https://arxiv.org/abs/2309.15505.

Mirzadeh, S.-I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., and Ghasemzadeh, H. Improved knowledge distillation via teacher assistant, 2019. URL https://arxiv.org/abs/1902.03393.

Miwa, K., Sasaki, K., Arai, H., Takahashi, T., and Yamaguchi, Y. One-d-piece: Image tokenizer meets quality-controllable compression, 2025. URL https://arxiv.org/abs/2501.10064.

Nash, C., Carreira, J., Walker, J., Barr, I., Jaegle, A., Malinowski, M., and Battaglia, P. Transframer: Arbitrary frame prediction with generative models. *arXiv preprint arXiv:2203.09494*, 2022.

Ni, Z., Wang, Y., Zhou, R., Guo, J., Hu, J., Liu, Z., Song, S., Yao, Y., and Huang, G. Revisiting non-autoregressive transformers for efficient image synthesis, 2024a. URL https://arxiv.org/abs/2406.05478.

Ni, Z., Wang, Y., Zhou, R., Han, Y., Guo, J., Liu, Z., Yao, Y., and Huang, G. Enat: Rethinking spatial-temporal interactions in token-based image synthesis, 2024b. URL https://arxiv.org/abs/2411.06959.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models, 2021a. URL https://arxiv.org/abs/2102.09672.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021b.

Park, Y.-H., Kwon, M., Choi, J., Jo, J., and Uh, Y. Understanding the latent space of diffusion models through the lens of riemannian geometry, 2023. URL https://arxiv.org/abs/2307.12868.

Peebles, W. and Xie, S. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Razavi, A., Van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Rippel, O., Gelbart, M., and Adams, R. Learning ordered representations with nested dropout. In *International Conference on Machine Learning*, pp. 1746–1754. PMLR, 2014.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

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, 2022.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models, 2022. URL https://arxiv.org/abs/2202.00512.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans, 2016. URL https://arxiv.org/abs/1606.03498.

Sargent, K., Hsu, K., Johnson, J., Fei-Fei, L., and Wu, J. Flow to the mode: Mode-seeking diffusion autoencoders for state-of-the-art image tokenization, 2025. URL https://arxiv.org/abs/2503.11056.

Sauer, A., Schwarz, K., and Geiger, A. Stylegan-xl: Scaling stylegan to large diverse datasets, 2022. URL https://arxiv.org/abs/2202.00273.

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. *arXiv preprint arXiv:2209.14792*, 2022.

Skorokhodov, I., Tulyakov, S., and Elhoseiny, M. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3626–3636, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models, 2023. URL https://arxiv.org/abs/2303.01469.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. URL https://arxiv.org/abs/1212.0402.

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., and Wilson, A. G. Does knowledge distillation really work?, 2021. URL https://arxiv.org/abs/2106.05945.

Tian, K., Jiang, Y., Yuan, Z., Peng, B., and Wang, L. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024. URL https://arxiv.org/abs/2404.02905.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Fvd: A new metric for video generation. 2019.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022.

Wang, Y., Ren, S., Lin, Z., Han, Y., Guo, H., Yang, Z., Zou, D., Feng, J., and Liu, X. Parallelized autoregressive visual generation, 2024. URL https://arxiv.org/abs/2412.15119.

Wang, Z., Jiang, Y., Zheng, H., Wang, P., He, P., Wang, Z., Chen, W., and Zhou, M. Patch diffusion: Faster and more data-efficient training of diffusion models, 2023. URL https://arxiv.org/abs/2304.12526.

Watson, D., Chan, W., Ho, J., and Norouzi, M. Learning fast samplers for diffusion models by differentiating through sample quality, 2022. URL https://arxiv.org/abs/2202.05830.

Weber, M., Yu, L., Yu, Q., Deng, X., Shen, X., Cremers, D., and Chen, L.-C. Maskbit: Embedding-free image generation via bit tokens, 2024. URL https://arxiv.org/abs/2409.16211.

Wen, X., Zhao, B., Elezi, I., Deng, J., and Qi, X. "principal components" enable a new language of images, 2025. URL https://arxiv.org/abs/2503.08685.

Yan, J. N., Gu, J., and Rush, A. M. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8239–8249, 2024.

Yang, S., Chen, Y., Wang, L., Liu, S., and Chen, Y. Denoising diffusion step-aware models, 2024. URL https://arxiv.org/abs/2310.03337.

Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A. G., Yang, M.-H., Hao, Y., Essa, I., et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023a.

Yu, L., Lezama, J., Gundavarapu, N. B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A. G., et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023b.

Yu, Q., He, J., Deng, X., Shen, X., and Chen, L.-C. Randomized autoregressive visual generation, 2024a. URL https://arxiv.org/abs/2411.00776.

Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., and Chen, L.-C. An image is worth 32 tokens for reconstruction and generation, 2024b. URL https://arxiv.org/abs/2406.07550.

Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.-W., and Shin, J. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022.

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., and Xie, S. Representation alignment for generation: Training diffusion transformers is easier than you think, 2024c. URL https://arxiv.org/abs/2410.06940.

Yuan, Z., Zhang, H., Lu, P., Ning, X., Zhang, L., Zhao, T., Yan, S., Dai, G., and Wang, Y. Ditfastattn: Attention compression for diffusion transformer models. *arXiv preprint arXiv:2406.08552*, 2024.

Zhao, Y., Xiong, Y., and Krähenbühl, P. Image and video tokenization with binary spherical quantization, 2024. URL https://arxiv.org/abs/2406.07548.

Zheng, C., Vuong, L. T., Cai, J., and Phung, D. Movq: Modulating quantized vectors for high-fidelity image generation, 2022. URL https://arxiv.org/abs/2209.09002.

Zheng, H., Nie, W., Vahdat, A., and Anandkumar, A. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

A. Motivation for Decode Time Model Scaling

Our visualization of token density averaged across 50k ImageNet samples reveals a dynamic pattern - initial decoding iterations prioritize background regions. In contrast, later iterations focus on the center where foreground objects or region of interest typically reside. This highlights the need to allocate resources efficiently during generation. To further investigate this behavior, we examine token density across various ImageNet categories (refer Figure 8). This category-wise analysis further motivates our focus on decode time scaling. Figure 10 shows more qualitative results on ImageNet256 × 256 and Figure 11 shows samples on UCF101.

B. Hyper-parameter Details

The MaskGIT algorithm has the following hyper-parameters which we discuss next.

Guidance Scale (gs). It is used in classifier-free guidance (Ho & Salimans, 2022) and governs the calculation of final logits during inference as shown in Equation (4).

$$\text{logits}_{\text{final}} = \text{logits}_{\text{cond}} + \lambda \cdot \text{gs} \cdot (\text{logits}_{\text{cond}} - \text{logits}_{\text{uncond}}) \tag{4}$$

where $\text{logits}_{\text{cond}}$ are from class-conditional input, $\text{logits}_{\text{uncond}}$ are from unconditional input, and λ depends on the mask-ratio of the current decoding iteration.

Figure 8 shows that the initial decoding iterations of parallel decoding focus on the background region, and focus gradually shifts to the main object/region in the final decoding iterations. Motivated by this, we experimented with applying guidance to only few final decoding iterations and present our findings in Figure 9b. As we can see, most of the decoding iterations do not require guidance. We use guidance only for final few decoding iterations for class-conditional generation in ImageNet256×256 and frame prediction in Kinetics600. Following MAGVIT (Yu et al., 2023a), for class-conditional generation in UCF101 we do not use classifier-free guidance.

Mask Temperature (MTemp). It controls the randomness introduced on top of the token predictions to mask tokens.

Sampling Temperature (STemp). It controls the randomness of the sampling from the categorical distribution of logits. Tokens are sampled from $\text{logits}/\text{STemp}$. STemp is calculated by Equation (5).

$$\text{STemp} = \text{bias} + \text{scale} \cdot (1 - (k+1)/K) \tag{5}$$

where bias and scale are hyperparameters (see Table 9), k is the current decoding iteration and K is the total number of decoding iterations. We report the hyperparameters we use in in Table 9. We use bias=0.5 and scale=0.8 for all experiments.

Dataset	Method	gs	MTemp
ImageNet	MaskGIT++	65	6
	MaGNeTS	65	5
UCF101	MAGVIT/ MaGNeTS	0	5
Kinetics600	MAGVIT	10	12.5
	MaGNeTS	5	10

Table 9: Best Sampling Hyperparameters.

C. Additional Experiments

C.1. Preliminary Diffusion experiments

We conduct initial experiments using model scheduling on diffusion models. Instead of training a new diffusion model with nesting and distillation, we focus solely on inference-time experiments. We use publicly available pretrained checkpoints of UViT (Bao et al., 2023) on ImageNet64×64. Specifically, we employ two models - U-ViT-L/4 (large) and U-ViT-M/4 (mid) - to investigate the impact of model scheduling during inference.

Implementation Details We use the default number of sampling steps of 50 and batch size of 500 in all experiments. We do not use classifier-free guidance. We do not use any caching for these experiments due to the continuous nature of the input. All experiments are run on a single A100 GPU.

Optimal Model Schedule Since the initial denoising steps play a crucial role in shaping the final output of the reverse diffusion process, we utilize the L model for these early stages and transition to the M model for the later denoising steps. Given that the L model has greater denoising capacity than the M model, we customize the noise schedule with larger denoising step sizes for L and smaller step sizes for M, balancing efficiency and performance.

Quantitative Results Refer Table 10 for results. With only model scheduling, we are able to achieve ~**1.53x** inference compute gains with almost similar performance as baseline. Exploring more refined schedules, training the models with nesting and distillation will offer better compute gains. Additionally, nesting would enable parameter sharing, unlike the current setup, which relies on separate models. This shows that the proposed method of model scheduling over multi-step decode process in image/video

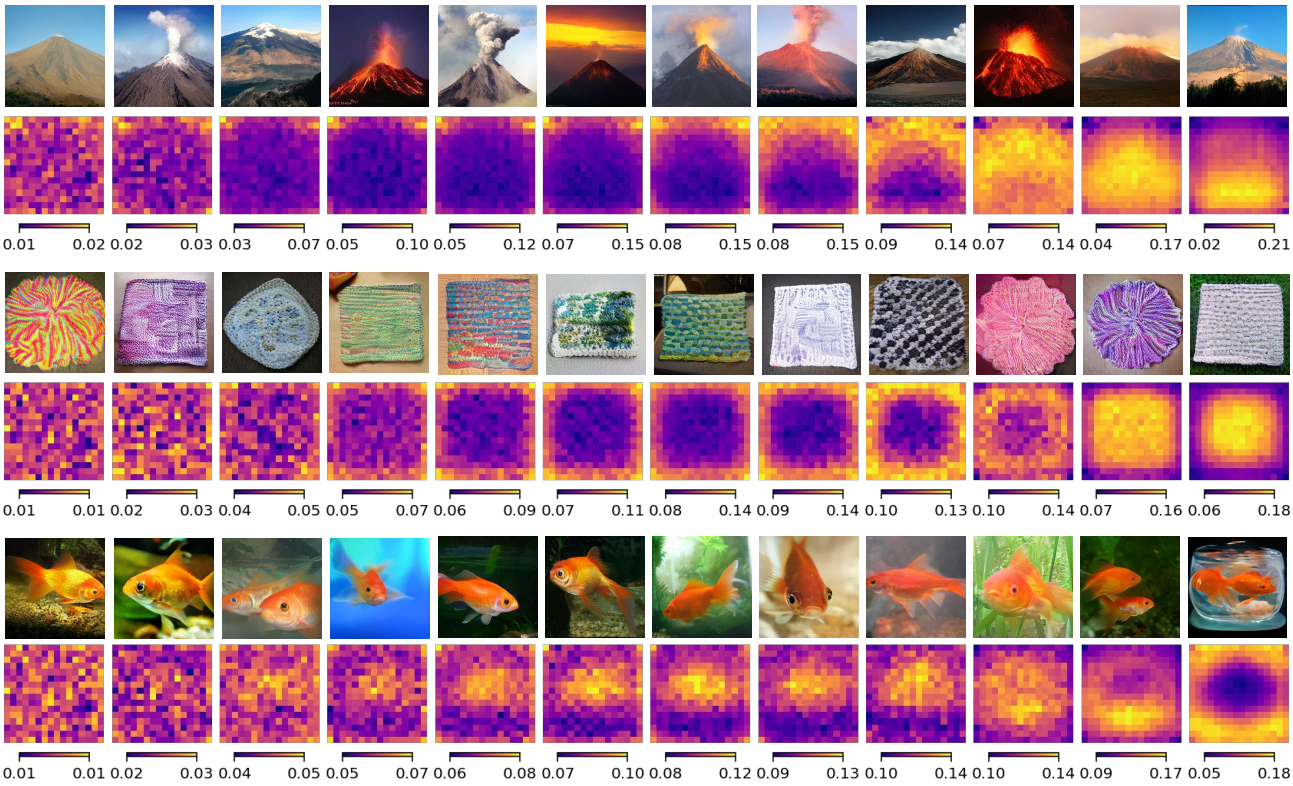


Figure 8: Visualization of token density unmasked in each iteration averaged over 10k generated samples on different categories of ImageNet. The top example shows category *volcano*. Middle and bottom examples show *"dishrag,dishcloth"* and *"goldfish,Carassius auratus"*, respectively. Yellow color represents higher density, and each pixel represents a token from the 16×16 token space.

generation is generic enough to be applied to different modeling approaches.

Method	FID (50k)	# params	# steps	Time (sec/iter)
U-ViT-M/4	5.92	131M	50	17.12
U-ViT-L/4	4.21	287M	50	32.34
Ours (model sched)	4.58	(131 + 287) M	50	21.10

Table 10: Class-conditional Image Generation on ImageNet64×64. "# steps" refers to the number of neural network runs.

D. Compute Gains

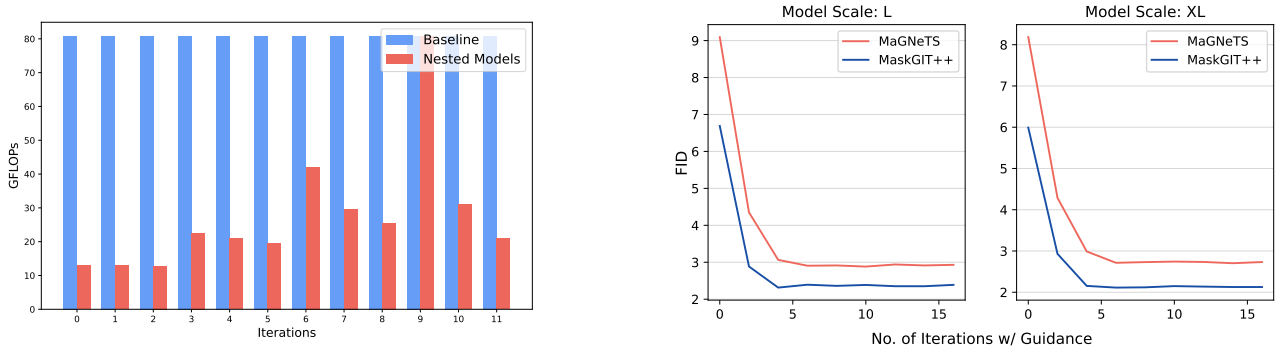
Per-step FLOPs. Figure 9a illustrates the inference-time computational cost, measured in GFLOPs, per iteration for the baseline model and MaGNeTS. As we can see the amount of FLOPs are drastically reduced using MaGNeTS. This is for a schedule with $k_1 = k_2 = k_3 = k_4 = 3$. The spikes after every 3 iterations are due to the cache refresh step. Mechanisms to get rid of the cache refresh can further reduce the total compute needed.

Calculation of GFLOPs. We illustrate the calculation of inference GFLOPs via Python pseudo-code in Table 12. We double the GFLOPs in decoding iterations where classifier-free guidance (Ho & Salimans, 2022) is used. Note that we always use a cosine schedule to determine the number of tokens to be unmasked in every step.

Real-Time Inference Benefits. In addition to the theoretical FLOP gains offered by MaGNeTS , here we want to analyze the real-time gains that it offers. We implement MaGNeTS on a single TPUv5 chip and present the results in Table 11.

Algorithm →	Baseline (MaskGIT++)	MaGNeTS
Images/Sec	22.5	56.3
Latency (ms)	712	285

Table 11: **Real-Time Inference Efficiency.** These show the number of generated images per sec and latency. These results are on ImageNet256×256 with model size XL.



(a) GFLOPs per sampling step.

(b) Guidance Analysis.

Figure 9: (a) Inference GFLOPs per step for baseline and MaGNeTS. (b) generation performance (FID) on ImageNet vs Number of decoding iterations w/ guidance for different model scales. Note that we start from last decoding iteration. For example, "No. of iterations w/ Guidance = 6" means we use guidance only for final six iterations (out of total 16 iterations). This shows that using guidance only for few final iterations is enough in the parallel decoding setup.

E. Limitations.

While our approach demonstrates strong performance in image and video generation, we acknowledge certain limitations. Some artifacts inherent to MaskGIT++ may also appear in our generated outputs (see Figure 12 for examples on ImageNet256×256). Such artifacts are common in models trained on controlled datasets like ImageNet. Moreover, the quality of the pretrained tokenizers (Yu et al., 2023b; Weber et al., 2024) directly impacts our method's effectiveness; however, improving these tokenizers is beyond the scope of this work. Although, use of nesting and decode time scaling does not have any specific requirement for model architecture and sampling scheme, to unlock the further benefits of KV caching, the process needs to generate discrete tokens.

Figure 10: **Class-conditional Image Generation.** More qualitative results on ImageNet. Comparing MaskGIT++ and MaGNeTS (size: L, epochs: 270).

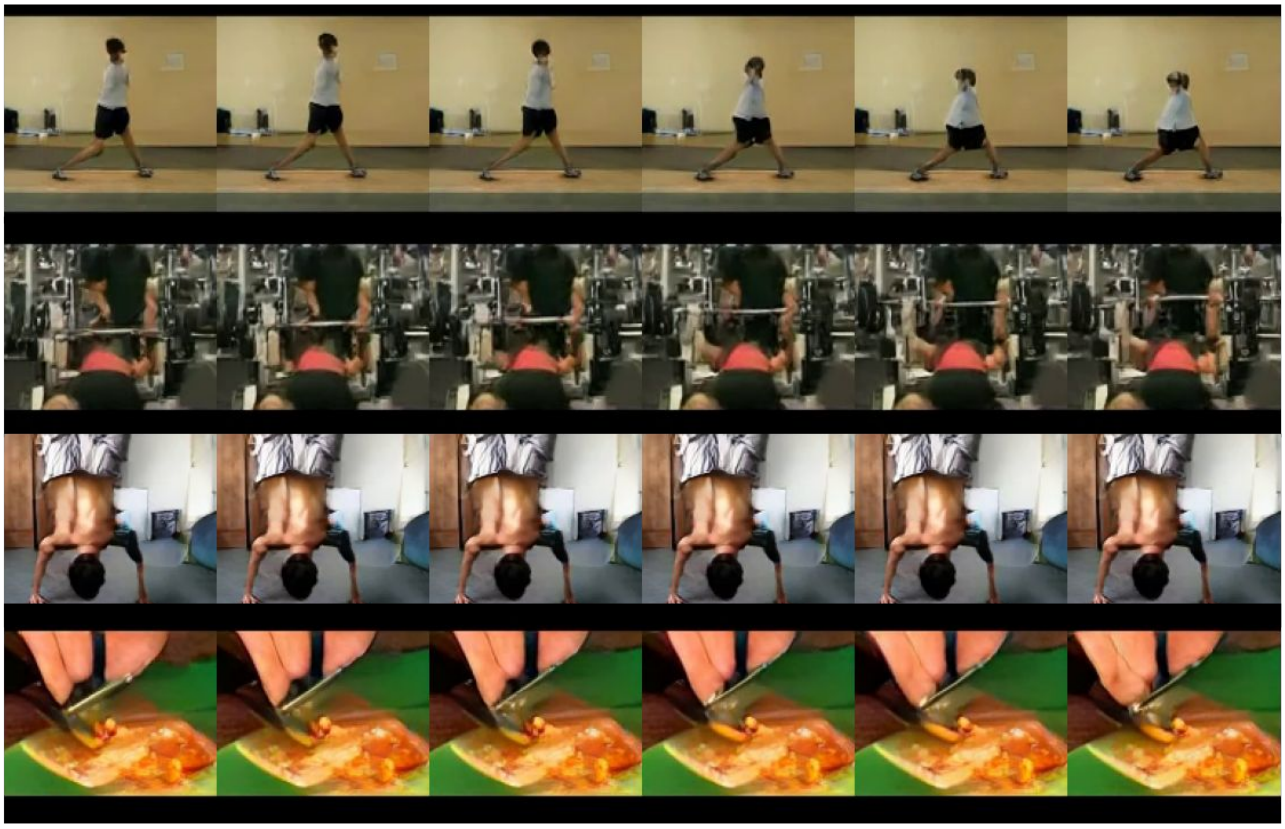


Figure 11: **Class-conditional Video Generation on UCF101.** 16-frame videos are generated at 128×128 resolution 25 fps. Every third frame is shown for each video. The classes from top to bottom are *Lunges*, *Bench Press*, *Handstand Pushups*, *Cutting In Kitchen*.

Figure 12: **Failure cases.** Similar to existing methods, our system can produce results with noticeable artifacts.

```python
# Function to get the GFlops for current decoding iteration
def get_flops(num_tokens_cached, num_tokens_processed, model_id, params, version):
    num_layers, hidden_size, mlp_dim, num_heads = params[version]
    qkv = 4 * num_tokens_processed * hidden_size * (hidden_size // model_id)
    attn = 2 * num_tokens_processed * (num_tokens_processed + num_tokens_cached) * hidden_size
    mlp = 2 * num_tokens_processed * (mlp_dim // model_id) * hidden_size
    return (qkv + attn + mlp) * num_layers // 1e9

# Function to get the total inference GFlops
def get_total_flops(version, num_iters, use_cache, refresh_cache_at, total_tokens, model_id_schedule, params, num_cond_tokens=0):
    assert num_cond_tokens < total_tokens
    refresh_cache_at = [int(x) for x in refresh_cache_at.split(',') if x]
    assert len(model_id_schedule) == num_iters
    num_cached = 0
    total_flops = 0

    # MaGNeTS (ours) doesn't need to process the conditioned tokens in the frame prediction task
    total_tokens -= num_cond_tokens

    for i in range(num_iters):
        ratio = i / num_iters

        # Cosine masking schedule
        num_processed = np.cos(np.pi/2. * ratio) * total_tokens

        # Even if we are performing caching, all tokens are processed in first iteration and iterations where cache is refreshed
        if i == 0 or i in refresh_cache_at and use_cache:
            total_flops += get_flops(0, total_tokens+num_cond_tokens, model_id_schedule[i], params, version)

        # we always cache the conditioned tokens
        else:
            total_flops += get_flops(num_cached+num_cond_tokens, total_tokens-num_cached, model_id_schedule[i], params, version)

        if use_cache:
            num_cached = total_tokens - num_processed
    return total_flops
```

```python
# Sample function call for class-conditional image generation
# params is a dictionary of the form {version: (num_layers, hidden_size, mlp_dim, num_heads)}
common = {'version': 'L', 'num_iters': 12, 'total_tokens': 257, 'params': params}
baseline = {'use_cache': False, 'refresh_at': '', 'model_id_schedule': (1,)*12, **common}
ours = {'use_cache': True, 'refresh_at': '3,6,9', 'model_id_schedule': (8,)*3+(4,)*3+(2,)*3+(1,)*3, **common}

print(get_total_flops(**baseline), get_total_flops(**ours))

# total_tokens = 1025 for class-conditional video generation and frame prediction
# num_cond_tokens = 512 for frame prediction
```

Table 12: **Python pseudo-code illustrating the calculation of inference GFLOPs.**