ACDiT: Interpolating Autoregressive Conditional Modeling and Diffusion Transformer

Anonymous authors
Paper under double-blind review

Abstract

The recent surge of interest in comprehensive multimodal models has necessitated the unification of diverse modalities. However, the unification suffers from disparate methodologies. Continuous visual generation necessitates the full-sequence-attend *diffusion*-based approach, despite its divergence from the *autoregressive modeling* in the text domain. In this paper, we explore an interpolation between autoregressive and modeling to model visual information. At its core, we present ACDiT, an **A**utoregressive blockwise **C**onditional **Di**ffusion **T**ransformer, where the block size of diffusion, i.e., the size of autoregressive units, can be flexibly adjusted to interpolate between token-wise autoregression and full-sequence diffusion denoising. ACDiT is easy to implement, as simple as creating a Skip-Causal Attention Mask (SCAM) during training. During inference, the process iterates between diffusion denoising and autoregressive decoding that can make full use of KV-Cache. We show that ACDiT performs best among all autoregressive baselines under similar model scales on image and video generation tasks. We also demonstrate that benefiting from autoregressive modeling, pretrained ACDiT can be transferred in visual understanding tasks despite being trained with the diffusion objective. The analysis of the trade-off between autoregressive modeling and diffusion demonstrates the potential of ACDiT to be used in long-horizon visual generation tasks. These strengths make it promising as the backbone of future unified models.

1 Introduction

The concept of predicting the future has been a fundamental principle of Artificial Intelligence, ranging from the "next token prediction" objective of autoregressive language modeling (Radford et al., 2019; Brown, 2020; Achiam et al., 2023) to predicting the next action in Reinforcement Learning (Mnih, 2013; Schulman et al., 2017; Chen et al., 2021; Reed et al., 2022). The recent renascence of the concept "World Model" also builds upon forming an internal state that is both capable of predicting the future and influenced by the prediction of the future (Ha & Schmidhuber, 2018). The success of Large Language Models (LLMs) (Touvron et al., 2023a;b; Brown, 2020; Achiam et al., 2023; Jiang et al., 2023; Bai et al., 2023; Yang et al., 2024) also exemplify the power of such philosophy, demonstrating that complex abilities can arise from this simple objective.

However, the multimodal aspects of information beyond language have yet to fully capitalize on this paradigm. In the realm of visual generation, diffusion models (Ho et al., 2020; Song et al., 2020a; Dhariwal & Nichol, 2021; Song et al., 2020b) have demonstrated superior generative capabilities, producing creative outputs that are virtually indistinguishable from human-generated content, as evidenced by innovations like Sora (Brooks et al., 2024) and Stable Diffusion (Rombach et al., 2022; Podell et al., 2023; Esser et al.). Notwithstanding their remarkable achievements, these models operate in a non-autoregressive manner. Diffusion models receive corrupted target sequences with complete length and reconstruct the intended output through *in-place* iterative refinement. The term "in-place" underscores the distinct nature of this approach compared to the autoregressive model, which provides subsequent prediction, thereby extending the sequence and progressing towards the future. Such variation makes the model struggle to learn temporal correlation in vision data and poses difficulties for developing integrated frameworks capable of seamlessly bridging the vision foundation model with unified multimodality modeling (Dong et al., 2024; Team, 2024; Zhou et al., 2024a; Wang et al., 2024b; Xie et al., 2024; Wu et al., 2024b) and world model (Yang et al., 2023; Du et al., 2023; Bruce et al., 2024; Zhou et al., 2024b; Ko et al., 2023; Wu et al., 2024a).

Existing works attempt to unify modalities by converting multimodal generation tasks into discrete token prediction tasks with vector quantization techniques (Esser et al., 2021; Team, 2024; Wang et al., 2024b; Tian et al., 2024) and training on the mixed sequences with a next-token prediction objective. However, approaching the continuous distribution requires huge vocabulary sizes and a high utilization rate (Yu et al., 2023; Weber et al., 2024), which is a complex objective. The information loss during vector quantization poses challenges for visual understanding tasks that require detailed information, such as the Optical Character Recognition task.

As a first step for exploring a unified framework, we propose ACDiT, an **A**utoregressive blockwise **C**onditional **Di**ffusion **T**ransformer that fuses the diffusion process with the autoregressive paradigm. At a high level, we extend the autoregressive units from the individual text token to blocks. The generation of each block can be formulated as a conditional diffusion process based on the previous block, where each block consists of visual patches of flexible size.

ACDiT is easy to implement, as simple as adding a Skip-Causal Attention Mask to the current DiT architecture during training. The inference process is formatted as an iteration between the conditional diffusion denoising process within a block, conditioned on the complete clean context, and autoregressive generation of a new block appended as the new context. In this way, KV-Cache can be used for faster inference. In general, ACDiT offers the following inherent advantages: (i) ACDiT simultaneously learns the causal interdependence across blocks with autoregressive modeling and the non-causal dependence within blocks with diffusion modeling. This serves as a versatile framework for expanding into unified multimodal and world models without conflict. (ii) ACDiT is endowed with clean continuous visual input, which can benefit visual understanding tasks in multimodal models. (iii) ACDiT makes full use of KV-Cache for flexible autoregressive generation in any length and potentially other latest long-context techniques in text for long video generation.

2 Related Work

Diffusion Models. The field of image generation has witnessed remarkable advancements with the introduction of diffusion models (Ho et al., 2020; Song et al., 2020a; Dhariwal & Nichol, 2021; Nichol & Dhariwal, 2021). U-Net (Ronneberger et al., 2015) is the early mainstream choice of network architecture (Song et al., 2020b; Nichol & Dhariwal, 2021; Rombach et al., 2022; Podell et al., 2023). Following that, Transformer (Vaswani et al., 2017) is applied to diffusion models for image generation, with groundbreaking work such as DiT (Peebles & Xie, 2022) and U-ViT (Bao et al., 2023) marking significant milestones. A series work, including PixArt-$\{\alpha, \delta, \Sigma\}$ (Chen et al., 2023; 2024c;b), demonstrate the capability of DiT on text-to-image tasks. Additionally, several studies have applied DiT to video generation, such as Lumiere (Bar-Tal et al., 2024) and Movie Gen (Polyak et al., 2024).

Autoregressive Visual Generation. Autoregressive models have shown promising results in visual generation (Chen et al., 2020; Esser et al., 2021; Tian et al., 2024; Li et al., 2024a; Wang et al., 2024b; Li et al., 2024b). The iGPT (Chen et al., 2020) first proposes autoregressively generating raw image pixels as a raster-scan sequence. VQGAN (Esser et al., 2021) improves the performance by training an autoregressive transformer on discrete tokens produced by VQVAE (Van Den Oord et al., 2017). LlamaGen (Sun et al., 2024) enhances the image tokenizer and scales up the autoregressive transformers to 3.1B parameters, building on the latest Llama architecture (Touvron et al., 2023a;b). Also, several works demonstrate the potential of token-base visual generation in text-to-image tasks (Yu et al., 2022; Liu et al., 2024; Sun et al., 2024; Wang et al., 2024b). Inspired by RQ-Transformer (Lee et al., 2022), VAR (Tian et al., 2024) proposes the next-scale prediction and obtain good improvement. In video generation, some works (Ho et al., 2022; Ruhe et al., 2024) utilize sliding windows for progressive generation. Diffusion Forcing (Chen et al., 2024a) and MAR (Li et al., 2024a) are two highly related works. Diffusion Forcing trains a causal autoregressive model to generate blocks without fully diffusing past ones and implement it on small RNN. MAR proposes the diffusion loss to learn the autoregressive conditional distribution on the head of the main Transformer with a small MLP network. Differently, ACDiT generates each block based on clear past and utilizes the full parameters of Transformer to denoise each block.

Unified Model for Understanding and Generation. Unified models for visual understanding and generation have recently garnered widespread attention. Some early efforts have aimed to align both the visual encoder and visual decoder with pre-trained Large Language Models (Dong et al., 2024;

Wu et al., 2023). Some works utilize the discrete visual token to unify the image understanding and generation tasks, such as VILA-U (Wu et al., 2024b), EMU3 (Wang et al., 2024b), and Show-o (Xie et al., 2024). Transfusion (Zhou et al., 2024a) tries the first attempt of joint training with language modeling loss and diffusion loss in a single transformer.

World Model. World models obtain great attention in various domains. Genie (Bruce et al., 2024) introduces interactive game video simulation by learning latent actions on unlabeled video. Unisim (Yang et al., 2023) builds real-world interaction simulations with web video data. iVideoGPT (Wu et al., 2024a) train an autoregressive transformer on a mixed sequence of tokens organized with visual observations, actions, and rewards. Despite the success, world models face challenges in leveraging the powerful generative capabilities of diffusion models due to their non-autoregressive nature.

3 PREREQUISTE

3.1 AUTOREGRESSIVE MODELING

Autoregression asserts that the value at each timestep is contingent upon its preceding values. This principle is exemplified in autoregressive language models, which iteratively predict the probability distribution of subsequent tokens. Given a sequence of tokens $(x_1, x_2, \ldots, x_n)$, a salient characteristic of autoregression is that the prediction of x_i is only dependent on its prefix $(x_1, x_2, \ldots, x_{i-1})$. Upon determining c_i, it is concatenated with the preceding sequence, thereby forming the conditioning context $(x_1, x_2, \ldots, x_i)$ for predicting x_{i+1}. Thus, the sequence likelihood can be factorized as:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_{i+1}|x_1, x_2, \ldots, x_i). \tag{1}$$

Thanks to the flexibility of self-attention in Transformers, autoregressive models can be effectively implemented by adding a causal mask in Transformer attention block (Vaswani et al., 2017).

3.2 DIFFUSION

Diffusion models, in contrast, conceptualize a noise-infusion and denoising process, which is defined by gradually adding noise to the initial data n_0 and training the model to learn the inverse mapping. Formally, the noised data $n^{(t)}$ at each step t is sampled by $q(n^{(t)}|n^{(t-1)}) = \mathcal{N}(x_t; \sqrt{\alpha^{(t)}}n^{(0)}, (1 - \alpha^{(t)})\mathbf{I})$[1], which is equivalent to add a Guassain noise to the previous samples: $n^{(t)} = \sqrt{\alpha^{(t)}}n^{(t-1)} + \sqrt{1 - \alpha^{(t)}}\epsilon^{(t)}, \epsilon^{(t)} \sim \mathcal{N}(0, \mathbf{I})$ while the diffusion model p_θ is trained to learn the reverse process $p_\theta(n^{(t-1)}|n^{(t)}) = \mathcal{N}(\mu_\theta(n^{(t)}), \beta^{(t)}\mathbf{I})$. With the reparameterization trick, the network $\mu_\theta(n^{(t)})$ can be reparameterized as noise prediction network $\epsilon_\theta(n^{(t)})$ and the training objective can be simple as:

$$\mathcal{L}_\theta = \mathbb{E}_{t\sim U[0,1],\epsilon\sim\mathcal{N}(0,I)}||\epsilon_\theta(n^{(t)}, t) - \epsilon^{(t)}||_2. \tag{2}$$

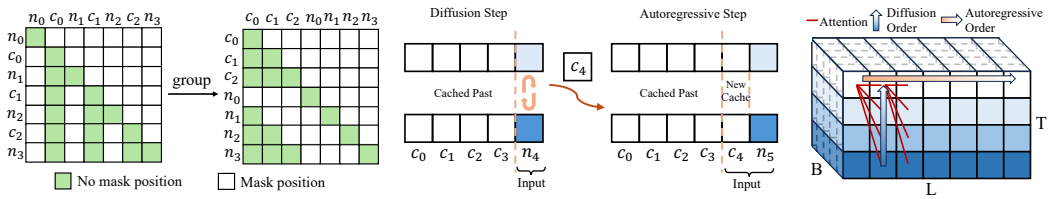
During the inference phase, the denoising process is initialized with a random Gaussian noise sample $n^{(T)}$, followed by T' denoising steps, ultimately yielding a single deterministic samples $\tilde{n}^{(0)}$ from its underlying distribution. Typically, the denoising process in diffusion models operates "in-place", meaning that each new denoising step directly replaces the previous step's input. This differs from autoregressive modeling, where the value of a subsequent step is appended to the existing sequence.

3.3 DESIDERATA

A robust autoregressive diffusion method should integrate the strengths of both autoregressive modeling and diffusion. To achieve this synergy, we have identified three critical desiderata that the framework must meet:

1. *The generation of future elements should be predicated on a precise representation of antecedent sequences.* This is imperative because any ambiguity in the past inevitably complicates future

[1]For clarification, we use subscript $_t$ to denote timesteps in autoregressive models and superscript $^{(t)}$ to denote the timesteps in diffusion models.



(a) Skip-Causal Attention Mask. (b) Inference process of ACDiT. (c) 3D view of ACDiT.

Figure 1: **(a)** For noised block n_i, it attends previous clean latent $c_0, c_1, \ldots,$ and c_{i-1}. Each clean block c_i only attends the previous clean latent block. **(b)** ACDiT effectively utilizes the KV-Cache for autoregressive inference. **(c)** The 3D view of ACDiT, where B, L, and T denote the block size, number of blocks, and denoising time steps, respectively. Darker color indicates higher noise levels.

predictions. This approach preserves the efficacy of autoregressive modeling and potentially facilitates the development of an "internal world model." Furthermore, adherence to this principle enhances performance in discriminative tasks (e.g., visual understanding), as these tasks necessitate the input of all observable features into the model.

2. *Both the autoregressive modeling and the denoising process should optimally utilize the entire parameter space of the neural network.* In an elegant fusion of autoregressive models and diffusion, neither component should be relegated to an auxiliary role. Instead, they should function as integral, complementary elements of the system.

3. *The denoising process should directly attend comprehensively to the entire sequence of past sequences.* Failure to do so would necessitate that the denoise process's condition encapsulates all preceding information, placing an unrealistic demand on the lossless compression capabilities of the feature space. On the contrary, a holistic input of all past information in each denoise step ensures a more effective processing of temporal dependencies.

Based on these desiderata, we analyze representative existing autoregressive diffusion methods. Diffusion-Forcing (Chen et al., 2024a) proposes to use different-level random noise in the different positions of a sequence. Thus, in the inference process, denoising the subsequent positions from a clean past and be seen as a special case denoising different levels of noise. Their method does not meet the first desideratum. In the training process, the future is not predicted from the precise representation of the past. In MAR (Li et al., 2024a), the diffusion process is trained based on the latent in the last position, which does not satisfy the third desideratum. Moreover, the diffusion process sorely leverages the head part of the network, which conflicts with the second desideratum. In Transfusion (Zhou et al., 2024a), despite that the diffusion utilizes the full network parameters for both autoregressive modeling and diffusion, it does not utilize the "predicting the future" objective within the multimodal information. Moreover, when trained with multiple blocks of multimodal information, such as multiple images, the latter image will attend to the noise version of the former image. To comprise this deficiency, they use half of the noise schedule, i.e., limiting maximum noise steps to 500 instead of 1000, in 20% image captioning pairs, which is not an optimal strategy.

4 ACDiT

4.1 Framework

To satisfy the desiderata discussed above, we propose a versatile framework for autoregressive diffusion called ACDiT. For generality, ACDiT runs block-wise autoregression instead of token-wise autoregression. We identify there are two kinds of blocks: the clean blocks c_i and the noise blocks n_i, where n_i is corrupted from c_i. ACDiT learns several conditional distribution $p(c_i|c_{<i})$ factorized with Eq. 1, where $p(c_i|c_{<i})$ is optimized by learning the conditional noise prediction network $\epsilon_\theta(n_i^{(t)}; t, c_{<i})$ with Eq. 2. The final training objective is:

$$\mathcal{L}_\theta = \mathbb{E}_{t\sim U[0,1],\epsilon\sim\mathcal{N}(0,I)} \sum_{i=1}^{n} ||\epsilon_\theta(n_i^{(t)}; t, c_{<i}) - \epsilon^{(t)}||_2. \tag{3}$$

Given the aforementioned desiderata, we can conceptualize all n_i and all c_i as occupying separate positions, effectively transforming the dependency structure into an attention pattern between different positions. We designate this pattern as the Skip-Causal Attention Mask (SCAM), which is shown in Figure 1a. The figure elucidates that n_i attends to all preceding clean blocks $\{c_j|j = 0, ..., i-1\}$ and itself, while c_i also attends to all preceding clean blocks $\{c_j|j = 0, ..., i-1\}$ and itself. In training, for both simplicity and efficiency, we can group the attention mechanisms as illustrated in the right matrix of Figure 1a. Suppose the number of blocks is N, then the unmasked positions form two triangular matrices of side length $N-1$, complemented by a diagonal matrix of side length N.

During the inference phase, each autoregressive step executes a conditional diffusion process for n_i based on $\{c_j|j = 0, ..., i-1\}$'s KV-Cache. Upon finishing denoising, it is appended to the clean sequence as c_i followed by the maximal noise version of the next block n_{i+1}. The key-value tensor will be computed for these two blocks, and the key-value tensor of the clean block c_i will be kept in KV-Cache. All noise-corrupted version of n_i is disregarded. The process is visualized in Figure 3. A three-dimensional view of our method is presented in Figure 1c. By bridging full-sequence diffusion and autoregressive paradigms, ACDiT gains flexibility and expressivity, allowing it to generate videos of any length using the latest long-context techniques from language models.

4.2 Positional Encoding

ACDiT is designed to be versatile, capable of handling one-, two-, three-, or even higher-dimensional data, including but not limited to text (1D), images (2D), and video (3D). For any given dimension of data, the position of that data is a critical attribute that must be made known to the model. This positional awareness enables the model to contextualize its current focus relative to historical data. In the domain of textual data, the Rotary Position Embedding (RoPE) (Su et al., 2024) has gained widespread adoption as an effective relative positional encoding method. To address the challenges posed by multi-dimensional positional indices, we introduce RoPE-ND, a natural extension of RoPE.

For a token of a D dimensional data, its positional index is $[m_1, m_2, ..., m_D]$. Given query and key vectors in Transformer's attention module, we partition the hidden dimension into D segments. It is imperative that each segment's hidden dimension be an even number. For each segment j, we apply a RoPE with a specific base b_j, as defined in the following equation:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}^{d_1}_{\Theta_1, m_1} & 0 & \cdots & 0 \\ 0 & \mathbf{R}^{d_2}_{\Theta_2, m_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{R}^{d_D}_{\Theta_D, m_D} \end{bmatrix} \tag{4}$$

In this formulation, each $\mathbf{R}^{d_j}_{\Theta_j, m_j}$ represents a d_j-dimensional rotary matrix[2] with rotation angles $\Theta_j = \{\theta_{ji} = b_j^{-2(i-1)/d_j}, i \in [1, 2, \cdots, \frac{d_j}{2}]\}$. The base b_j is empirically determined as $100\lceil\frac{8L_j}{100\pi}\rceil$, where L_j denotes the maximum position index in data dimension j. This formulation ensures that the highest wavelength of RoPE is approximately eight times the maximum position, thereby mitigating rapid decay in long-term dependencies. It is worth noting that ACDiT inherently supports length extrapolation (Su et al., 2024), although a comprehensive exploration of this capability falls beyond the scope of the present work.

4.3 Efficiency Analysis and Block Size Choice

In this section, we provide a brief analysis on the computational efficiency of ACDiT in terms of floating-point operations (FLOPS). We first assume that denoising each block requires the same number of time steps T as the full sequence diffusion, despite that when the block size is small, it may require fewer denoising steps, thus making ACDiT potentially more efficient. Let θ denote the number of parameters in one layer of the transformer block, h represents the hidden dimension and n as the number of attention heads. The FLOPS is formed by a $O(L)$ component $2L\theta$, from the product between input and weight matrices, and a $O(L^2)$ component from the pairwise computation between

[2]For a detailed explanation, please refer to Equation 15 in version 5 of the arXiv preprint of (**?**)su2024roformer.

Q-K pairs in the attention dot product, which is $4(h+n)$ for each Q-K pair in the L^2 Q-K pairs. Consequently, given that $n \ll h$, the total FLOPS can be expressed as: $F = 2L\theta + 4(h+n)L^2 \approx 2L\theta + 4hL^2$.

When employing ACDiT with KV-Cache, the FLOPS consumed by FFN and attention parameters remain unchanged. However, the attention mechanism transforms into causal attention with block size B. In this scenario, the number of Q-K pairs to be calculated is given by:

$$\sum_{i=1}^{L/B} iB^2 = \frac{1}{2}(\frac{L}{B} + \frac{L^2}{B^2})B^2. \tag{5}$$

Thus, the FLOPS saved can be expressed as $4hL^2(\frac{1}{2}(1-\frac{B}{L}))$. The saved percentage is

$$\frac{4hL^2(\frac{1}{2}-\frac{B}{2L})}{4hL^2+2L\theta} = \frac{1-B/L}{2+m/k}, \tag{6}$$

where $m = \frac{\theta}{h^2}$ and $k = \frac{L}{h}$. This equation shows that when the sequence length significantly exceeds the hidden size, transforming a full sequence diffusion into ACDiT can save up to 50% of FLOPS.

However, in practice, it is not always beneficial for small B. Setting B to an excessively small value may not fully leverage the iterative modification inherent in the diffusion process, potentially compromising generation quality. Furthermore, given the parallel computing nature of computational kernels, a very small B may not yield speed improvements, a phenomenon analogous to the rationale behind speculative decoding (Leviathan et al., 2023). Conversely, setting B to a very large value diminishes efficiency both in terms of attention calculation, as shown in Equation 6. It also fails to capitalize on the strengths of auto-regressive generation. We analyze the influence of block size on performance and efficiency in Sec. 5.3.

4.4 Model Architecture

ACDiT mainly inherits the main architecture of DiT. However, since we want to keep the architecture as simple and unified as possible, we use linear layers instead of convolution in the input layer and final layer. Besides, we replace the absolute position embedding and Layer Normalization with RoPE (Su et al., 2024) and RMSNorm (Root Mean Square Layer Normalization) (Touvron et al., 2023a), respectively. We find that QK-norm is important to stabilizing the video generation training, thus we use QK-norms in all experiments. The additional conditional information timesteps and labels are injected into the model with adaLN-Zero only on the noise part. For both image and video generation, we follow DiT and leverage the pre-trained image VAE (Kingma, 2013) from Stable Diffusion (Rombach et al., 2022), whose downsample factor is 8. For image generation under B block size, we group square latent representation patches with $\sqrt{B} \times \sqrt{B}$ shape as a block.

5 Experiments

5.1 Experimental Setup

Dataset. For image generation tasks, we consider the ImageNet (Russakovsky et al., 2015) dataset with 256×256 resolution, which consists of around 1.28M images from 1K classes. For video generation, we consider the UCF-101 (Soomro et al., 2012) dataset with 16 frames, where each frame is an image with 256×256 resolution. UCF-101 contains 13320 videos from 101 classes.

Implementation details. In the image generation task, we set the patch size as 1 and the autoregressive unit block size as $256 = 16 \times 16$. Therefore, for a $256 \times 256 \times 3$ image in $32 \times 32 \times 4$ latent shape, the total sequence length and autoregressive length are 1024 and 4, respectively. We explore 4 different model sizes, as shown in Table 4. ACDiT-B is used for design verification and analyze. ACDiT is trained on ImageNet for 1.2M iterations with a batch size of 1024. We use the AdamW optimizer (Loshchilov, 2017) and WSD (Warmup Steady Decay) learning rate scheduling (Hu et al., 2024) with the peak learning rate 3e-4 and no weight decay. The learning rate begins to decay in the last 15% training iteration. Following the common training recipe of generative models, we keep

Table 1: Image generation results on ImageNet 256x256.

Model	Type	Latent	KV-Cache	Params	FID↓	IS↑	Pre↑	Rec↑
ADM (Dhariwal & Nichol, 2021)	Diff.	-	-	554M	10.94	101.0	0.69	0.63
LDM-4-G (Rombach et al., 2022)	Diff.	Cont.	-	400M	3.60	247.7	-	-
DiT-XL/2 (Peebles & Xie, 2022)	Diff.	Cont.	-	675M	2.27	278.2	0.83	0.57
MaskGIT (Weber et al., 2024)	Mask.	Disc.	-	227M	6.18	316.2	0.83	0.58
MAGE (Li et al., 2023)	Mask.	Disc.	-	230M	6.93	195.8	-	-
VQGAN (Esser et al., 2021)	AR	Disc.	✓	1.4B	15.78	78.3	-	-
RQTran (Lee et al., 2022)	AR	Disc.	✓	3.8B	7.55	134.0	-	-
VAR-d16 (Tian et al., 2024)	VAR	Disc.	✓	310M	3.30	274.4	**0.84**	0.51
VAR-d20 (Tian et al., 2024)	VAR	Disc.	✓	600M	2.57	**302.6**	0.83	0.56
LlamaGen-L (Sun et al., 2024)	AR	Disc.	✓	343M	3.07	256.1	0.83	0.52
LlamaGen-XL (Sun et al., 2024)	AR	Disc.	✓	775M	2.62	244.1	0.80	0.57
LlamaGen-XXL (Sun et al., 2024)	AR	Disc.	✓	1.4B	2.34	253.9	0.80	0.59
ImageFolder (Li et al., 2024b)	AR	Disc.	✓	362M	2.60	295.0	0.75	**0.63**
MAR-L (Tian et al., 2024)	AR	Cont.	✓	479M	4.07	232.4	-	-
MAR-L (Tian et al., 2024)	MAR	Cont.	-	479M	**1.78**	296.0	0.81	0.60
ACDiT-L	AR+Diff	Cont.	✓	460M	2.53	262.9	0.82	0.55
ACDiT-XL	AR+Diff	Cont.	✓	677M	2.45	267.4	0.82	0.57
ACDiT-H	AR+Diff	Cont.	✓	954M	2.37	273.3	0.82	0.57

Table 2: Video generation results on UCF-101. ACDiT-XL-LT means training for longer epoch.

Model	Type	Params	FVD↓
LVDM (He et al., 2022b)	Diff.	437M	372
Latte (Ma et al., 2024)	Diff.	674M	478
Matten (Gao et al., 2024)	Diff.	853M	211
VideoFusion (Luo et al., 2023)	Diff.	510M	173
MMVG (Fu et al., 2023)	Mask.	230M	328
MAGVITv2 (Yu et al., 2023)	Mask.	307M	**58**
TATS (Ge et al., 2022)	AR	331M	332
CogVideo (Hong et al., 2022)	AR	9.4B	626
MAGVITv2-AR (Yu et al., 2023)	AR	307M	109
OmniTokenizer (Wang et al., 2024a)	AR	650M	191
ACDiT-XL	AR+Diff.	677M	111
ACDiT-H	AR+Diff.	954M	104
ACDiT-H-LT	AR+Diff.	954M	<u>90</u>

Table 3: Supervised fine-tuned Top-1 accuracy on Imagenet.

Model	Type	Top-1 Acc
ViT-H	Supervised	83.1
MAGE	Masked.	84.3
MAE	Masked.	85.9
iGPT	Generative	72.6
DiT-XL	Generative	82.8
ACDiT-XL	Generative	84.0

an exponential moving average (EMA) of the ACDiT weights during training using a decay rate of 0.9999. We sample images with DPM-Solver (Lu et al., 2022) for 25 steps within each block and use classifier-free guidance (Ho & Salimans, 2022) with a guidance scale of 1.5. In video generation, we sample 16 frames from each video and set the patch size as 2 and the block size as $1024 = 256 \times 4$. For a $16 \times 256 \times 256 \times 3$ video in $16 \times 32 \times 32 \times 4$ latent shape, the sequence length of each frame is 256 and the total sequence length is 4096, with 4 frames grouped into one block. We train ACDiT on UCF-101 for 400K iterations with a batch size of 96. The classifier-free guidance scale is 2.5. Other training configs are the same as image training. All models are implemented with PyTorch (Paszke et al., 2019) and trained on NVIDIA H100 GPUs. Specifically, we use FlexAttention[3] to implement the SCAM for both customization and efficiency.

5.2 Main Results

Image Generation. We report the FID-50K (Heusel et al., 2017), Inception Score (Salimans et al., 2016), Precision and Recall (Kynkäänniemi et al., 2019) of ACDiT and baselines in Table 1. Compared with previous autoregressive models and masked generative models utilizing discrete tokens, such as VQGAN, VAR, LlamaGen, and MaskGIT, ACDiT consistently achieves superior performance with lower FID scores at comparable model scales. Notably, ACDiT-XL achieves 2.45 FID scores, outperforming both LlamaGen-XXL and VAR-d20 with similar parameters. Additionally,

[3]https://pytorch.org/blog/flexattention

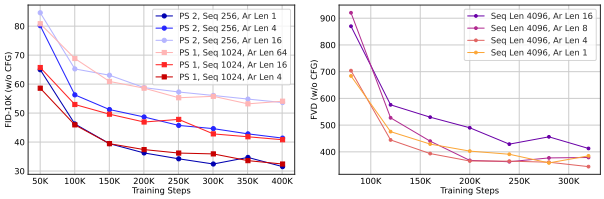


Figure 2: FID and FVD curves of ACDiT-B over training steps with different sequence lengths and autoregressive lengths. PS means patch size.

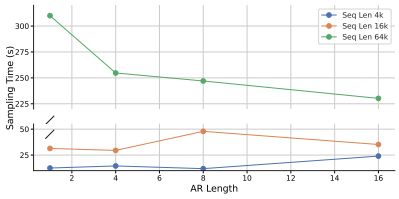
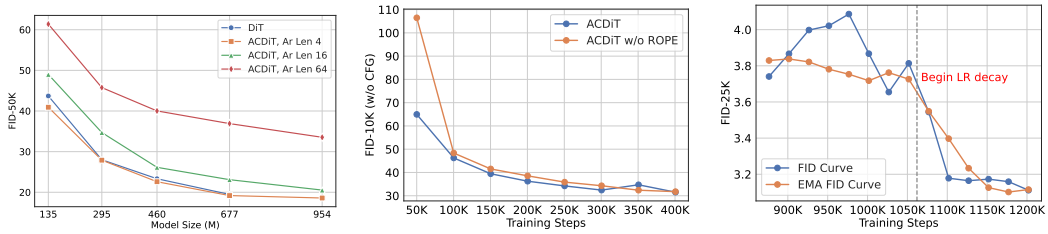


Figure 3: The change of inference time in different autoregressive lengths under varied total sequence length.



(a) Scaling performance of ACDiT. (b) Ablation study for ROPE-ND. (c) FID curve of last 30% training.

Figure 4: **(a)**: ACDiT shows scaling performance similar to DiT. **(b)**: ROPE-ND has consistent improvement to the generation quality. **(c)**: FID score sharply with the learning rate beginning to decay when using the WSD scheduler.

when compared to the MAR-L variant that does not recompute attention, ACDiT-L significantly improves performance across all metrics. Although MAR-L has lower FID than ACDiT, recomputing attention makes it hard to generalize longer sequence generation. When compared with leading diffusion-based methods, ACDiT also demonstrates competitive performance. For instance, despite not employing full-sequence attention, ACDiT models achieve results close to DiT-XL. In general, these results highlight the distinct advantages of ACDiT over other baselines with the continuous latent representation and KV-Cache. Qualitative results are presented in Fig. 5.

Video Generation. Different from image generation, video inherently includes a temporal dimension, making it more well-suited to autoregressive modeling. The FVD metric on UCF-101 for class-conditional video generation is reported in Table 2. With hybrid AR+Diff architecture, ACDiT-H achieves much lower FVD than other diffusion-based and autoregressive methods, even outperforming MAGVITv2-AR, which utilizes a closed-source, specially designed video tokenizer. In contrast, ACDiT simplifies the process by directly using an open-sourced image VAE. Although MAGVITv2 with masked generative methods has a lower than ACDiT, they rely on "in-place" operation to generate a video similar to the diffusion model. This constraint limits their ability to generalize to generate longer video generation and build world models. Compared to image generation, ACDiT demonstrates greater potential in modeling long visual sequences. Qualitative results of ACDiT-XL are presented in Fig. 6.

Image Representation. We also assess the capability of ACDiT in image representation, which is essential for building a unified visual understanding and generation model. We finetune ACDiT-XL and DiT-XL on ImageNet using classification loss following the training setting in MAE (He et al., 2022a) and report the Top-1 accuracy in Table 3. The accuracy of ACDiT surpasses that of ViT-H, iGPT and DiT-XL. This superior performance over DiT-XL highlights the benefit of incorporating clean latent inputs, which accelerates the model's ability to learn better representations compared to using only noised latent inputs. Furthermore, ACDiT is on par with MAGE in terms of Top-1 accuracy, while enjoying better generation capabilities than MAGE.

5.3 Analysis

Trade off of block size. Fig. 2 illustrates the trend of trade-off under different sequence lengths and block sizes in image and video generation tasks on ACDiT-B. The FID curve indicates that for image generation, directly increasing the autoregressive length leads to a decline in image quality, since each patch receives less attention information on average. However, we can mitigate this decline

Figure 5: Sample images from ACDiT-H on ImageNet 256×256.

Figure 6: Sample videos from ACDiT-XL trained on UCF-101.

by increasing the total sequence length, which means reducing the patch size. For video generation, ACDiT shows more advantages due to the inherent temporal dependence of videos. The FVD curve demonstrates that increasing autoregressive length has minimal effect on the video quality, even with slight improvement. As for efficiency, we test the sampling time for various sequence lengths with a batch size of 4 on an NVIDIA A100 GPU. Fig. 3 shows that as the sequence length increases, particularly beyond 16k, full-sequence attention (AR length of 1) becomes very time-consuming, necessitating the autoregressive generation.

Scaling Performance. We present the scaling performance of ACDiT in Fig. 4a. For a fair comparison with DiT, we use the same batch size and learning rate as DiT in these training sessions. When increasing the model size, ACDiT shows consistent improvement in image quality across all autoregressive lengths, sharing a similar scaling trend with DiT. Notably, the improvement is more pronounced with longer autoregressive lengths. We hypothesize that this is due to reduced accumulation of errors when scaling the model size.

Ablation Study. We ablate the effectiveness of ROPE-ND positional embedding on image generation with patch size as 2. As shown in Fig 4b, adding ROPE-ND results in consistent improvements.

Training dynamics of WSD scheduler. Unlike the constant learning rate used in DiT, we utilize the WSD learning rate scheduler (Hu et al., 2024). WSD scheduler maintains a constant learning rate as the main stage of training, while one can diverge from the main branch at any time, potentially based on the compute budget, with a rapidly decaying learning rate. As Fig. 4c shows, the FID remains almost converged during the constant learning rate state, while sharply dropping after the learning decays, similar to the loss curve when using the WSD scheduler in LLM training. To the best of our knowledge, we are the first to validate the effectiveness of the WSD scheduler in visual generation.

6 Conclusion

In this paper, we propose ACDiT that interpolates the autoregressive modeling and diffusion transformers. With a simple but novel design of attention mask, ACDiT can achieve autoregressive generation on any length while maintaining a clear latent input potentially for adding a visual understanding task. We demonstrate the performance and efficiency of ACDiT in image and video generation tasks while endowing sufficient generalization by combining the advantages of both autoregressive and diffusion models. We hope ACDiT can shed light on the architectural design of building a unified multimodal model and world model in the future.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024b.

Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-{\delta}: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024c.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020.

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

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=y01KGvd9Bw.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10681–10692, 2023.

Yu Gao, Jiancheng Huang, Xiaopeng Sun, Zequn Jie, Yujie Zhong, and Lin Ma. Matten: Video generation with mamba-attention. *arXiv preprint arXiv:2405.03025*, 2024.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pp. 102–118. Springer, 2022.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022a.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. *arXiv preprint arXiv:2310.08576*, 2023.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024a.

Xiang Li, Hao Chen, Kai Qiu, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024b.

Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Volodymyr Mnih. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012. URL https://api.semanticscholar.org/CorpusID:7197134.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *arXiv preprint arXiv:2406.09399*, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. Maskbit: Embedding-free image generation via bit tokens. *arXiv preprint arXiv:2409.16211*, 2024.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models, 2024a.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2023.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024a.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024b.

Table 4: Configuration details of ACDiT.

Model	#Layers	Hidden Size	MLP	#Heads	Params
ACDiT-B	12	768	3072	12	132M
ACDiT-L	24	1024	4096	16	460M
ACDiT-XL	28	1152	4608	18	677M
ACDiT-H	32	1280	5120	20	954M

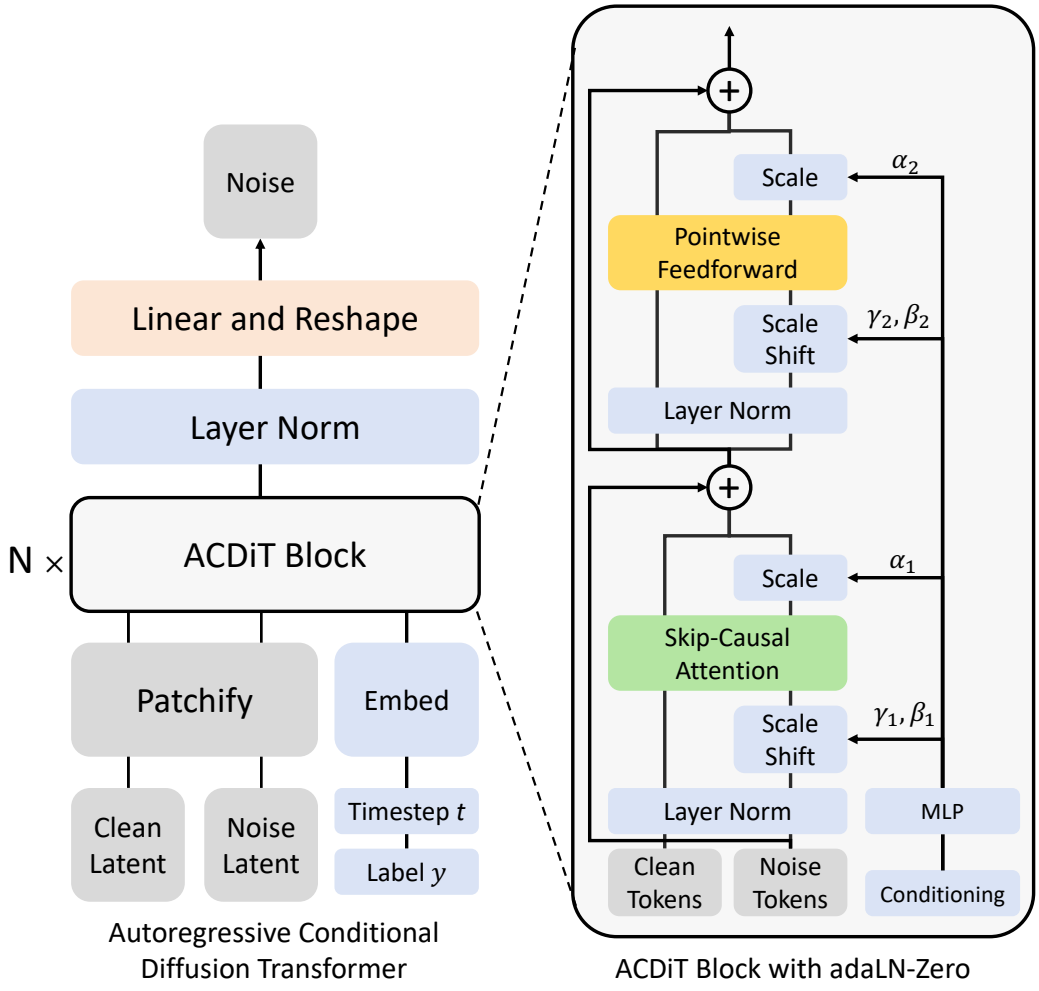


Figure 7: The model architecture of ACDiT. Both the clean latent and noise latent are input while only noise tokens are scale shift with conditioning information.