ELT: Elastic Looped Transformers for Visual Generation

Anonymous CVPR submission

Paper ID *****

Abstract

We introduce Elastic Looped Transformers (ELT), a highly parameter-efficient class of visual generative models based on a recurrent transformer architecture. While conventional generative models rely on deep stacks of unique transformer layers, our approach employs iterative, weight-shared transformer blocks to drastically reduce parameter counts while maintaining high synthesis quality. To effectively train these models for image and video generation, we propose the idea of Intra-Loop Self Distillation (ILSD), where student configurations (intermediate loops) are distilled from the teacher configuration (maximum training loops) to ensure consistency across the model's depth in a single training step. Our framework yields a family of elastic models from a single training run, enabling Any-Time inference capability with dynamic trade-offs between computational cost and generation quality, with the same parameter count. ELT significantly shifts the efficiency frontier for visual synthesis. With $4\times$ *reduction in parameter count under iso-inference-compute settings, ELT achieves a competitive FID of* 2.0 *on class-conditional ImageNet* 256×256 *and FVD of* 72.8 *on class-conditional UCF-101.*

1. Introduction

Traditional techniques to increase compute capacity in deep learning models, such as stacking deeper layers or increasing network width, inevitably lead to a proportionally larger memory footprint. Recurrence offers a powerful alternative, decoupling compute from memory via parameter reuse. While popularized by Universal Transformers [2] and recently leveraged for language reasoning [24, 31], looping remains largely untapped for high-fidelity visual generation. Compared to standard deep models, Looped Transformers offer three distinct practical advantages: (a) **extreme parameter efficiency** (more FLOPs per parameter), (b) have **higher throughput** by minimizing the "memory wall" bottleneck. They use a compact set of shared parameters and maintain its major parameter footprint on-chip or adjacent to the chip. This avoids the cost of repeated transfers between different units of the accelerator (GPUs/TPUs) required in large transformers, and (c) can exhibit robustness against overfitting in data-constrained regimes.

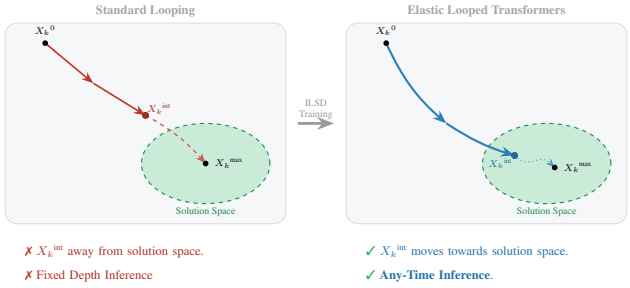
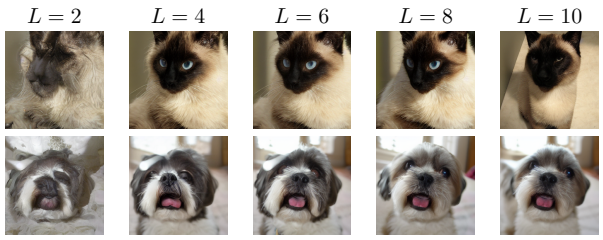


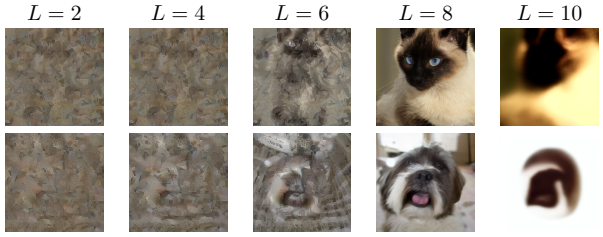
Figure 1. **Latent Trajectories of Standard vs. Elastic Looped Transformers**. $X_k{}^{\text{int}}$ & $X_k{}^{\max}$ represent output of intermediate (L_{int}) & final loops ($L_{\max}$) respectively for k^{th} generation sampling step. Unlike standard recurrent models **(left)** where only the final iteration $X_k{}^{\max}$ reaches the solution space, our ILSD training **(right)** guides intermediate states $X_k{}^{\text{int}}$ also toward the target space. This transformation shifts the model from a fixed-depth architecture to an Any-Time inference framework, supporting flexible computational budgets through early exits within a sampling step.

Despite the parameter efficiency of looped architectures, their training remains challenging because intermediate representations often remain uninterpretable until the final loop (see Fig. 1). We address this by introducing Elastic Looped Transformers (ELT), a class of generative models designed for progressive refinement. Unlike traditional recurrent transformers, ELT provides meaningful, high-quality synthesis even at intermediate repeats, enabling Any-Time (elastic) inference - where a single model can scale its compute based on available resources without sacrificing generation quality (Figs. 2 and 3).

To achieve this flexibility, we propose an Intra-Loop Self Distillation (ILSD) algorithm for training looped transformers. Rather than treating the loop as a fixed-depth process, our framework operates as a dual-path system: a teacher path executes the full loop count to provide a high-fidelity target, while a student path, defined strictly as a subset of the teacher's trajectory, learns to produce comparable re-



(a) Any-Time inference from a single training run



(b) Vanilla looped transformers

Figure 2. **Class-conditional Image Generation on ImageNet 256 × 256**. Comparing ELT based diffusion model with ILSD (left) and without ILSD (right). ELT is trained with $N = 4$ and $L_{\text{max}} = 8$. Number of loops during inference (L) is varied from 2 to 10. ELT with ILSD (left) offers good generation quality for a variety of loops, which is not the case for the vanilla looped transformers.

sults with fewer iterations. Since the student computation is a literal subset of the teacher's forward propagation, there is no additional training overhead. This forces the shared parameters to compress complex transformations into earlier loops, learning a progressive refinement process that maintains generation quality regardless of exit point (Fig. 1). Our contributions and findings are summarized as follows:

State-of-the-Art Parameter Efficiency: ELT achieves a competitive FID of *2.0* on class-conditional ImageNet 256× 256 and FVD of *72.8* on class-conditional UCF-101. This represents a *4*× reduction in parameters compared to baselines MaskGIT [1] (image generation) and MAGVIT [32] (video generation), while matching or improving performance under iso-inference-compute settings.

Elastic/Any-Time Inference: Our models enable Any-Time inference [35], traversing the pareto frontier of quality versus compute at test-time without retraining.

Scalability: Recursive looping provides a unique test-time compute lever that scales predictably across both Masked Generative Transformers [1, 32, 33] and Diffusion Transformers [19].

2. Background

Masked Generative Transformers: MaskGIT [1] introduced iterative parallel decoding for image generation using discrete tokens [32]. Unlike autoregressive models, MaskGIT generates and refines all tokens simultaneously over K sampling steps:

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

where M is a fixed-depth transformer predicting tokens conditioned on class c. MAGVIT [32] extends this framework to video generation. **Diffusion Transformers:** Diffusion models [11, 25] generate data by reversing a noising process. Diffusion Transformers (DiTs) [19] shift away from traditionally used U-Nets [22] by treating image latents as sequences of tokens, and using transformer blocks for processing these tokens.

Both paradigms share a common structure: recursive refinement over multiple sampling steps through a model M. ELT naturally aligns with this progressive refinement by implementing M as a recurrent, weight-shared transformer block. This introduces recursive computation *within* each sampling step, providing a test-time compute lever to dynamically trade-off inference speed and generation quality.

3. Method

Looping Mechanism: Let the number of transformer layers to be looped be N and number of loops per sampling step be L, giving an effective depth of $N \times L$. Let $f_{\theta_i}(\mathbf{x})$ denote a single transformer layer with parameters θ_i. We define a composite block $g_\Theta(\mathbf{x}) = f_{\theta_N}(f_{\theta_{N-1}}(\cdots f_{\theta_1}(\mathbf{x})))$ with $\Theta = \{\theta_1, \ldots, \theta_N\}$. In *looping*, we reuse Θ for L successive applications, requiring only N unique layers instead of $N \times L$ sets of unique parameters:

$$F_{(N,L)}(\mathbf{x}) = \underbrace{g_\Theta\left(g_\Theta(\cdots g_\Theta(\mathbf{x}))\right)}_{L \text{ loops}} \equiv g_\Theta^L(\mathbf{x})$$

This decouples physical model size from computational depth: the parameter count is constrained by N, while representational capacity scales with L. Note that we represent the training loop count by L_{max}.

Intra-Loop Self Distillation (ILSD): In standard looped transformers, the model is optimized only for its final output after L_{max} iterations, creating a "black box" trajectory where intermediate loops may produce suboptimal representations as shown in Fig. 2. We propose Intra-Loop Self Distillation (ILSD) to ensure the model remains useful at multiple depths, enabling elastic inference with competitive performance. By treating the full-depth model (L_{max}) as an internal teacher, we provide a high-fidelity signal for the shallower intermediate student (L_{int}), forcing Θ to compress complex transformations into fewer steps.

Training: We train with a fixed L_{max} loops. At each training step, we randomly sample an intermediate loop count $L_{\text{int}} \sim \mathcal{U}(L_{\text{min}}, L_{\text{max}})$. The teacher executes the full L_{max}

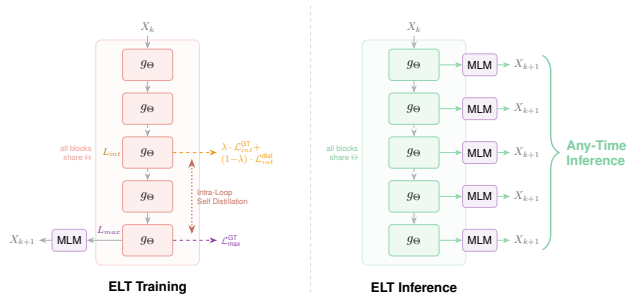


Figure 3. **Overview of ELT framework**. Training (left) utilizes shared parameters Θ with recurrent loops and Intra-Loop Self Distillation (ILSD) to improve intermediate representations. This enables Any-Time inference (right), allowing the model to exit early and predict X_{k+1} from any intermediate block via a shared MLM head. Note that X_k represents the input for k^{th} sampling iteration.

loops while the student exits early at L_{int}. The joint loss is:

$$\begin{aligned}\mathcal{L}_{\Theta}^{\text{ILSD}} &= \mathcal{L}^{\text{GT}}(F_{(N,L_{\text{max}})}(\mathbf{x}), \mathbf{y}) \\ &+ \lambda\, \mathcal{L}^{\text{GT}}(F_{(N,L_{\text{int}})}(\mathbf{x}), \mathbf{y}) \\ &+ (1-\lambda)\, \mathcal{L}^{\text{dist}}(F_{(N,L_{\text{int}})}(\mathbf{x}), F_{(N,L_{\text{max}})}(\mathbf{x})) \end{aligned} \quad (2)$$

The first term supervises the teacher with ground-truth $\mathbf{y}$, the second supervises the student with ground-truth, and the third distills the teacher into the student. λ is linearly decayed from 1 to 0 during training—initially anchoring the student to reliable ground-truth labels while the teacher is untrained, then gradually shifting to mimicking the teacher once it has matured. For masked generative models, $\mathcal{L}^{\text{GT}}$ and $\mathcal{L}^{\text{dist}}$ are cross-entropy losses over masked positions; for diffusion models, they are sigmoid-weighted MSE losses [13].

Since both paths share Θ, the joint optimization forces the block to learn a transformation that is effective at any depth $L_{\text{int}} \leq L \leq L_{\text{max}}$, preventing shortcuts that minimize loss at a specific depth but fail when composed further.

4. Experiments and Results

We evaluate ELT on class-conditional image generation using both masked generative transformers and diffusion transformers, and on class-conditional video generation using masked generative transformers.

4.1. Experimental Setup

Datasets: ImageNet 256×256 [3] for images; UCF-101 [27] for videos. **Masked Generative Transformers**: We use pretrained tokenizers from MaskGIT [1] (images) and MAGVIT [32] (videos) with a codebook size of 1024. Images are compressed to 16×16 tokens; videos to $4 \times 16 \times 16$ tokens. We adopt BERT [5] as the backbone, training all models for 270 epochs with classifier-free guidance (10% class-condition drop). **Diffusion Transformers**:

Table 1. **Class-conditional Image Generation** on ImageNet 256×256. "# steps" refers to the number of neural network runs. $^{\Delta}$ denotes values taken from prior publications. * indicates usage of extra training data. ^{g} denotes use of classifier-free guidance [10]. Note that L in $(N \times L)$ notation is inference loop count per sampling step.

Model	AR	FID ↓	IS ↑	# params	# steps	# Gflops
ADM + Upsample$^{\Delta g}$ [6]		3.9	215.8	554M	250	371k
LDM-4$^{\Delta g*}$ [20]		3.6	247.7	400M	250	51.5k
DiT-XL/2$^{\Delta g*}$ [19]		2.3	278.2	675M	250	59.5k
MDT$^{\Delta g*}$ [7]		1.8	283.0	676M	250	>59k
MaskDiT$^{\Delta g*}$ [34]		2.3	276.6	736M	250	>28k
RIN$^{\Delta}$ [14]		3.4	182.0	410M	1000	334k
Simple Diffusion$^{\Delta g}$ [12]		2.4	256.3	2B	512	-
VDM++$^{\Delta g}$ [15]		2.1	267.7	2B	512	-
EDiff$^{\Delta g}$ [8]		2.1	-	450M	50	119k
MAR$^{\Delta g}$ [16]	✓	1.8	296.0	479M	128	-
MaskGIT$^{\Delta}$[1]		6.2	182.1	227M	8	647
MaskBit$^{\Delta g}$ [30]		1.7	341.8	305M	64	10.3k
PAR-4×$^{\Delta}$ [29]	✓	3.8	218.9	343M	147	-
PAR-16×$^{\Delta}$ [29]	✓	2.9	262.5	3.1B	51	-
MaskGIT-L^{g}		2.1	270.1	303M	24	3.7k
MaskGIT-XL^{g}		2.0	294.8	446M	24	3.9k
ELT-L ($8N \times 3L$)		2.2	254.3	**101M**	24	3.7k
ELT-L ($12N \times 2L$)		2.1	281.8	**152M**	24	3.7k
ELT-XL ($7N \times 4L$)		2.0	266.1	**111M**	24	3.9k

We use a pretrained Stable Diffusion VAE [21] to map images into $32 \times 32 \times 4$ latents and train a DiT [19] with sigmoid-weighted MSE loss [13] for 500K steps. Sampling uses 512-step DDPM with guidance scale 3.0. **Metrics**: FID [9], IS [23] for images; FVD [28] for videos. Efficiency is measured via inference GFLOPs and throughput.

4.2. Image Generation

Comparison with Baselines: Tab. 1 presents results on ImageNet. Despite using $\mathbf{4}\times$ fewer parameters, ELT-XL achieves the same FID of 2.0 as MaskGIT-XL baseline. Using superior tokenizers [30, 33] or optimized training & inference configurations [17, 18] can further boost ELT's performance. We also evaluate ELT framework for diffusion models in Tab. 2. Notably, ELT with $8N \times 4L$ outperforms the DiT-16 and DiT-32 baselines, achieving $\mathbf{2}\times$ and $\mathbf{4}\times$ parameter reduction respectively. While looping without ILSD gives competitive performance when running inference with same loops as training ($L = L_{\text{max}}$), performance degrades drastically for lower number of loops which is mitigated by ILSD as show in Fig. 6b.

Scaling and Throughput: Fig. 4 illustrates the trade-off between generation quality (FID) and inference compute (GFLOPs). The pareto front (black curve) reveals that while increasing the loop count (L) consistently improves FID (faded points) for a fixed unique layers (N), the gains eventually diminish, where transitioning to the next architecture scale becomes more performant than over-looping smaller models. Crucially, ELT allows for Any-Time inference, we can traverse the pareto curve at test-time by simply adjusting L to meet specific hardware constraints without retraining. Moreover, ELT has high throughput as it utilizes a compact set of shared parameters and maintains its major

Table 2. **ImageNet (DiT)**. Inference with $N \times L$.

Model	FID ↓	# params
DiT - 16 layers	3.87	1.1B
DiT - 32 layers	3.43	2.1B
ELT ($1N \times 32L$)	10.30	69M
ELT ($4N \times 8L$)	3.96	271M
ELT ($8N \times 4L$)	**3.16**	**539M**
ELT ($16N \times 2L$)	**2.83**	**1.1B**

Table 3. **Throughput Gains**. ELT vs. Baseline.

ELT	d_{model}	Ratio
$6N \times 2L$ (B)	768	1.0
$8N \times 3L$ (L)	1024	2.9
$7N \times 4L$ (XL)	1152	3.3
$8N \times 4L$ (H)	1280	**3.5**

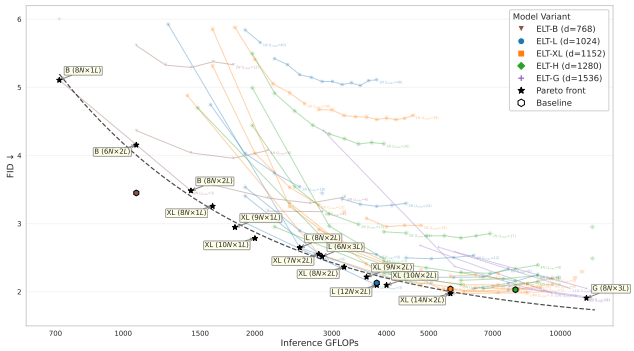


Figure 4. **Pareto front of FID vs. Inference GFLOPs**. The black curve denotes the fit (FID = $1922.5 \cdot G^{-0.95} + 1.48$) over pareto-optimal configurations, representing the best achievable FID for a given computational budget. Points are labeled as N layers $\times$ L loops. Results demonstrate that ELT scales as effectively as the baseline while remaining significantly more parameter-efficient. Scaling both the model dimension (d) and the number of loops (L) follows a predictable efficiency frontier, where larger models with fewer loops often compete with smaller models with more loops at specific target GFLOPs. Trends across faded points show scaling along number of loops L in inference, from a single run trained with a fixed L_{max} loops.

weight footprint closer to the accelerator computation unit. This avoids the cost of repeated memory transfers typically required in standard models, achieving a peak speedup of *3.5*$\times$ throughput gains for model scale H (refer Tab. 3). These gains scale with model size as long as shared parameters fit within device memory.

4.3. Video Generation

We extend ELT to class-conditional video generation on UCF-101 using the MAGVIT [32] framework. As shown in Tab. 4, our compact 76M ELT model outperforms the MAGVIT baseline in iso-compute settings on UCF-101 ($\sim$13.7M training tokens), achieving FVD of 72.8 with 4$\times$ fewer parameters. This suggests that looped transformers exhibit robustness against overfitting in data-constrained regimes like UCF-101.

Table 4. **UCF-101 Video**. ELT vs. Baselines.

Method	FVD ↓	# params	GFlops
TATS$^{\Delta}$	332	321M	-
Make-A-Video$^{\Delta *}$	81	$\gg$3.5B	-
PAR-4$\times^{\Delta}$	99.5	792M	-
MaGNeTS$^{\Delta}$	96.4	306M	$\sim$1.7k
MAGVIT-L$^{\Delta}$	76	306M	$\sim$4.3k
ELT ($6N \times 4L$)	72.8	**76M**	$\sim$4.3k
ELT ($6N \times 6L$)	**60.8**	**76M**	$\sim$13k

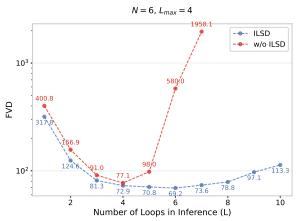
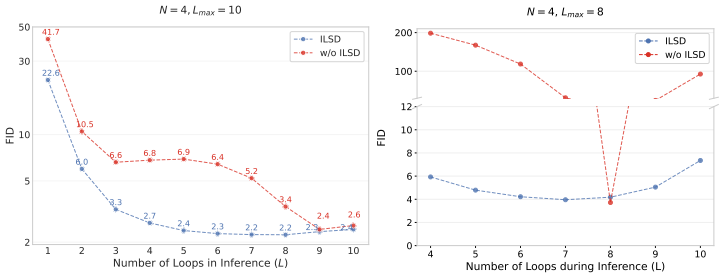


Figure 5. **Impact of ILSD** on class-conditional UCF-101.



(a) Masked Generative Models (b) Diffusion Transformers

Figure 6. **Impact of ILSD.** ILSD improves performance for all L in inference especially when $L \neq L_{max}$. Results on ImageNet.

4.4. ILSD drives Elasticity

Fig. 6 analyzes ILSD's impact across masked generative models and diffusion models. Models trained without ILSD degrade significantly at $L \neq L_{\text{max}}$, while ELT maintains stable generation quality across the entire loop spectrum. Remarkably, as shown in Fig. 5, ILSD enables zero-shot generalization to unseen depths for UCF-101 generation, reaching peak FVD of 69.2 at $L = 6$ despite training with $L_{\text{max}} = 4$.

5. Conclusion

We proposed Elastic Looped Transformers (ELT), a parameter-efficient visual generation framework using recurrent transformers. Through Intra-Loop Self Distillation (ILSD), ELT achieves performance comparable to baselines with 4$\times$ fewer parameters while enabling Any-Time inference to dynamically balance quality and compute at test-time. Looking forward, applying ELT to one-step generative paradigms, such as consistency [26] or drifting models [4], could unlock true elasticity. Since there is only one sampling step, one can dynamically control the quality of the model at inference by varying the number of loops, without having to pre-determine the number of sampling steps as is the case with traditional diffusion models. Ultimately, this flexible, weight-efficient scaling offers a promising path for deploying high-fidelity generative models on resource-constrained hardware.

References

[1] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

[2] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

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

[7] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23164–23173, 2023.

[8] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy, 2024.

[9] 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.

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

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

[12] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.

[13] Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 fid on imagenet512 with pixel-space diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18062–18071, 2025.

[14] Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. *arXiv preprint arXiv:2212.11972*, 2022.

[15] Diederik P Kingma and Ruiqi Gao. Understanding the diffusion objective as a weighted integral of elbos. *arXiv preprint arXiv:2303.00848*, 2, 2023.

[16] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization, 2024.

[17] Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, and Gao Huang. Revisiting non-autoregressive transformers for efficient image synthesis, 2024.

[18] Zanlin Ni, Yulin Wang, Renping Zhou, Yizeng Han, Jiayi Guo, Zhiyuan Liu, Yuan Yao, and Gao Huang. Enat: Rethinking spatial-temporal interactions in token-based image synthesis, 2024.

[19] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[21] 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*, pages 10684–10695, 2022.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[23] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[24] Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers, 2025.

[25] 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*, 2020.

[26] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023.

[27] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

[28] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[29] Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, and Xihui Liu. Parallelized autoregressive visual generation, 2024.

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

[31] Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. *arXiv preprint arXiv:2311.12424*, 2023.

[32] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023.

[33] 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.

[34] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

[35] Shlomo Zilberstein. Using anytime algorithms in intelligent systems. *AI magazine*, 17(3):73–73, 1996.