

One Model, Many Budgets: Elastic Latent Interfaces for Diffusion Transformers

Moayed Haji-Ali^{1,2,†} Willi Menapace² Ivan Skorokhodov² Dogyun Park^{2,†} Anil Kag²
 Michael Vasilkovsky² Sergey Tulyakov² Vicente Ordonez¹ Aliaksandr Siarohin²
¹Rice University ²Snap Inc.

Project Webpage: <https://snap-research.github.io/ELIT>

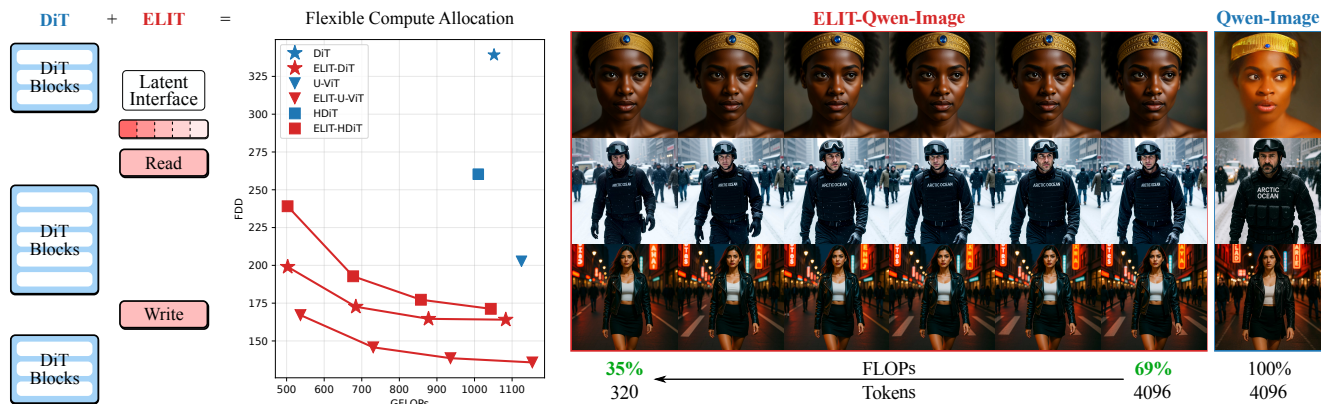


Figure 1. **Flexible compute allocation with ELIT.** Starting from a vanilla DiT, we add a variable-length set of latent tokens—the *latent interface*—and two lightweight cross-attention layers, Read and Write. At inference, the number of latent tokens is a user-controlled knob that yields a smooth quality–FLOPs trade-off across DiT, U-ViT, HDiT, and MM-DiT backbones.

Abstract

Diffusion transformers (DiTs) achieve high generative quality but lock FLOPs to image resolution, limiting principled latency-quality trade-offs, and allocate computation uniformly across input spatial tokens, wasting resource allocation to unimportant regions. We introduce Elastic Latent Interface Transformer (ELIT), a drop-in, DiT-compatible mechanism that decouples input image size from compute. Our approach inserts a latent interface, a learnable variable-length token sequence on which standard transformer blocks can operate. Lightweight Read and Write cross-attention layers move information between spatial tokens and latents and prioritize important input regions. By training with random dropping of tail latents, ELIT learns to produce importance-ordered representations with earlier latents capturing global structure while later ones contain information to refine details. At inference, the number of latents can be dynamically adjusted to match compute constraints. ELIT is deliberately minimal, adding two cross-attention layers while leaving the rectified flow

objective and the DiT stack unchanged. Across datasets and architectures (DiT, U-ViT, HDiT, MM-DiT), ELIT delivers consistent gains. On ImageNet-1K 512px, ELIT delivers an average gain of 35.3% and 39.6% in FID and FDD scores.

1. Introduction

Recent years have seen dramatic progress in image and video generation. Owing to the simplicity of their design, architectures centered on Diffusion Transformers (DiTs) [45] have scaled reliably and delivered state-of-the-art fidelity [19, 54, 56]. Compute has been the primary determinant of generation quality, but continued scaling has inflated training and inference costs. This raises a central question: do DiTs utilize available computation effectively? We argue that such costs are amplified by the *rigidity* in the DiT design. First, a DiT typically commits to a per-step computational cost that is a fixed function of the input resolution, without accounting for latency and compute constraints. Second, we found that DiT allocates computation uniformly across image regions. In a controlled experiment, we probe the ability of a DiT to use extra compute to improve generation quality. As expected, quality

[†]Work partially done during an internship at Snap Inc.

improves on standard images when we increase the number of tokens by reducing the patchification size. However, when we increase the number of tokens by padding encoded image patches with zero-valued patches, we find that DiT fails to leverage the extra computation to improve generation quality. These observations suggest that compute is spent *uniformly* across image tokens. This is suboptimal since visual information in images is uneven: some regions are easy, others require more work. In this context, *learning* how to allocate computation across tokens through a flexible architecture holds the potential to dynamically reduce unnecessary computation.

Important previous work has sought forms of flexibility to make generative models more efficient. Adaptive generators allow a *single* model to operate at different budgets, but still spread computation uniformly across tokens [1, 62] or suffer from high complexity [64]. Masking-based methods improve training speed by skipping tokens. Yet, dropping is disabled for inference to avoid unrecoverable information loss [17, 33], making inference compute unchanged. On the other hand, some training-free acceleration approaches reduce the computation for the least informative tokens during inference, leaving the training efficacy the same [28, 41, 59]. A complementary thread moves flexibility in the autoencoder by learning variable-length representations but stops short of endowing a *generative* model with an internal flexible representation [2, 36]. Finally, RINs [12, 26] learn to distribute computation non-uniformly across input tokens through a set of latent tokens, but keep inference budget fixed and depart significantly from the DiT architecture, hindering adoption.

Building on the previous observations, we propose Elastic Latent Interface Transformer (ELIT) (see Figure 1), a minimal, DiT-compatible mechanism for representation-and-compute adaptivity. We introduce two lightweight cross-attention layers, *Read* and *Write*, that equip DiT-like architectures with a set of latent tokens we refer to as *latent interface*. These latent tokens act as a variable-size surface onto which to distribute input information in a flexible and learnable manner based on the difficulty of each region. *Read* pulls information from input tokens, which we will refer to as *spatial*, into the latent interface, prioritizing challenging regions. *Write* broadcasts updated latents back to the spatial tokens. Importantly, the number of latent tokens is user-controlled, directly setting the per-step compute budget. No changes to the training objective are necessary.

We provide a thorough analysis of ELIT, showing that it successfully redistributes compute non-uniformly across input tokens across varying base architectures. Our latent interface consistently improves over a fixed-grid model with ImageNet-1k 512px FDD (Fréchet Distance on DINOv2 [44] features), improving 58.0% for DiT, 34.0% for U-ViT, 37.4% for HDiT. ELIT also allows for graceful

compute-quality tradeoffs by selecting the number of latent tokens at inference time, regularly achieving better tradeoffs than sampling steps reduction while being compatible with training-free acceleration techniques [39]. Additionally, variable compute enables autoguidance [30] out of the box, which reduces inference cost by $\approx 33\%$ without affecting the generation quality. In summary, this simple addition yields a framework capable of: (I) **Adaptive computation**. Compute is concentrated where it matters rather than spread uniformly across input regions. (II) **Variable test-time compute**. A single set of weights serves a spectrum of latency-quality points by selecting the number of latent tokens. (III) **Improved sampling**. Variable compute enables autoguidance [30] out of the box. (IV) **Drop-in training**. We keep the design simple, retaining the vanilla rectified-flow objective and showing our method’s applicability to DiT, U-ViT, HDiT, and MMDiT. Implementation amounts to only adding the Read and Write layers and latent tokens sampling during training.

2. Related Work

Adaptive generators for inference budget control. Supernetwork trains a single set of weights to support many sub-networks, allowing test-time accuracy–efficiency tradeoffs [5, 62]. Other works train transformers with multiple patchification sizes for variable compute budgets [1, 38]. [64] uses learnable routers to adjust network width and drop tokens in MLPs. These methods differ in *where* adaptivity lives but share the goal of a *single* model that gracefully scales compute at inference time. Our method adopts a simple variable-length latent interface, improving model convergence and enabling control over the inference budget. **Token dropping for training speedup.** Another direction accelerates model training by skipping tokens. MaskDiT [65] restructures DiTs as an encoder-decoder model and randomly drops encoder input while using an auxiliary reconstruction objective. MDTv2 [17] similarly leverages masked latent modeling to train on partial inputs. TREAD [33] randomly selects a set of tokens that skip computation of all blocks from a predefined start to an end DiT block index. However, due to the destructive nature of token dropping, such methods typically rely on auxiliary losses [65], full-token post-training [33, 65], and adopt full-token inference [17, 33, 65], limiting acceleration during inference. Our method speeds up convergence while enabling control over inference budget by selecting variable amounts of tokens, without applying auxiliary losses.

Latent interfaces. Latent tokens have been used as compact representations in several architectures. Neural Turing Machines [18] employed them as memory, while Perceiver [27] used cross-attention to condense high-dimensional inputs. In generative models, RINs [26] and FITs [12] introduced interleaved read/write operations for

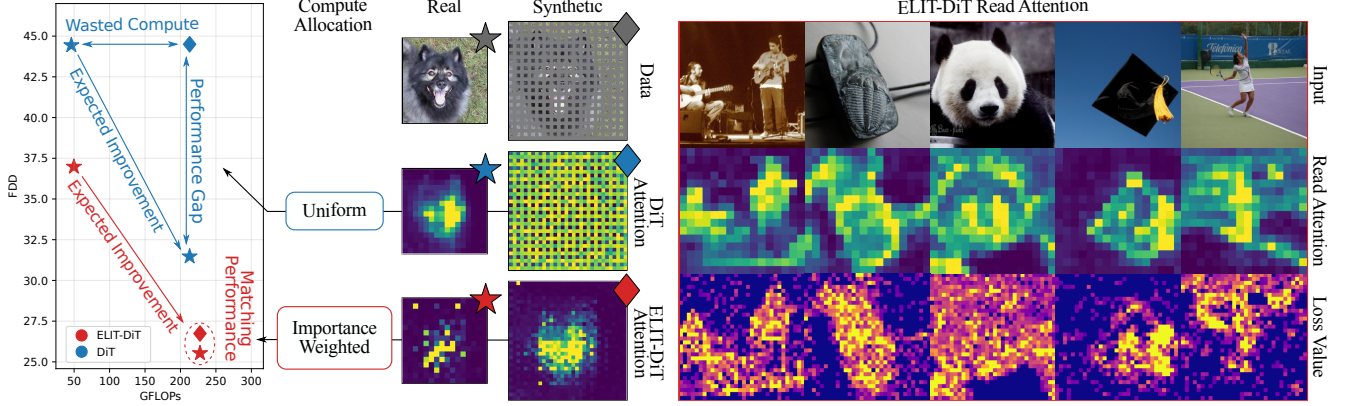


Figure 2. **Adaptive computation.** We test whether DiT and ELIT-DiT can reallocate compute across image regions by training on synthetic inputs formed by zero-padding real images, artificially increasing the token count (◆). We compare its performance to baselines trained on real data using patch size 2×2 and patch size 1×1 (★). Vanilla DiT does not improve: attention in zeroed regions targets other zeroed regions (see “DiT Attention”), so extra tokens raise cost without benefits. In contrast, ELIT-DiT uses the Read layer to pull informative spatial tokens into the latent interface (see “Read Attention”), effectively filtering out zeroed areas (see “ELIT-DiT Attention”). Consequently, it leverages the larger token budget and matches the real-data baseline at equal FLOPs.

efficient high-dimensional synthesis, which was later scaled to video generation [43]. Despite their efficiency, such designs diverge from DiTs and often require specialized optimizers (e.g., LAMB [61]). Similarly, TiTok [63] applied latent tokens as bottlenecks in autoencoders, and recent work extends this to variable-length token sets via tail dropping [2, 32, 36, 58]. Our work brings variable-length latent interfaces to generative models, integrating seamlessly into DiTs with only lightweight Read and Write layers.

3. Method

We propose Elastic Latent Interface Transformer (ELIT) (see Figure 3), a novel approach that enables flexible compute allocation in DiT-like transformers. The core component is a variable-length set of latent tokens—the *latent interface*—where most transformer blocks operate. Two lightweight cross-attention layers, Read and Write, pass information between domains: Read pulls content from spatial tokens into the latent interface, prioritizing harder regions, while Write broadcasts the updated latent state back to the spatial domain. Unlike the spatial domain, where model FLOPs are a fixed function of input resolution, the latent interface is trained with random tail token dropping, making it resizable. The length of this latent interface during inference directly adjusts FLOPs for each model call.

Section 3.1 reviews Flow Matching, Section 3.2 presents a motivating experiment, Section 3.3 details the architecture, and Section 3.4 describes training and sampling.

3.1. Preliminaries: Flow Matching

We train our generators with the Rectified Flow (RF) variant of flow matching [37, 40], which learns a deterministic velocity field connecting a noise distribution p_n to the data

distribution p_d . Let $\mathbf{X}_1 \sim p_d$ and $\mathbf{X}_0 \sim p_n = \mathcal{N}(0, \mathbf{I})$. A linear path is defined as $\mathbf{X}_t = (1-t)\mathbf{X}_0 + t\mathbf{X}_1$, $t \in [0, 1]$, whose ground-truth velocity is constant along the path: $v_t = \frac{d\mathbf{X}_t}{dt} = \mathbf{X}_1 - \mathbf{X}_0$. A neural network $\mathcal{G}(\cdot)$ predicts the velocity from a noised input and time and is optimized as:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{t \sim p_t, \mathbf{x}_1 \sim p_d, \mathbf{x}_0 \sim p_n} \|\mathcal{G}(\mathbf{x}_t, t) - (\mathbf{X}_1 - \mathbf{X}_0)\|_2^2, \quad (1)$$

where p_t is a logit-normal [15] training distribution over t . At inference, samples are obtained by integrating the learned ODE from $\mathbf{X}_0$ to $\mathbf{X}_1$ with a standard solver.

3.2. Uniform Spatial Computation in DiTs

Standard DiTs operate in the spatial domain where an input $\mathbf{X}_t$ at time t is patchified by a linear projection layer into N tokens and processed by B transformer blocks. Each block output is connected to the previous block through a *residual connection* at the same spatial location, maintaining a fixed mapping between tokens s in intermediate blocks and corresponding spatial location of $\mathbf{X}_t$. This results in a uniform compute distribution across spatial locations in $\mathbf{X}_t$.

We probe this behavior through an experiment presented in Figure 2. We train DiT-B/2 (i.e., 2×2 patchification) and a corresponding DiT-B/1 (i.e., 1×1 patchification) on ImageNet-1k. As expected, performance improves in DiT-B/1 due to the fourfold amount of tokens. We then introduce a synthetic ImageNet-1k version by padding encoded images with zeros, yielding four times more tokens. We train a DiT-B/2, named DiT-B/2-Synth, without loss on padded regions. We also use absolute learnable positional encodings to avoid bias toward zero patches. For evaluation, we remove padded regions before decoding to recover real image content. DiT-B/2-Synth matches the number of tokens and FLOPs of DiT-B/1. Thus, if compute were used effectively, it should match DiT-B/1 performance. Instead, as

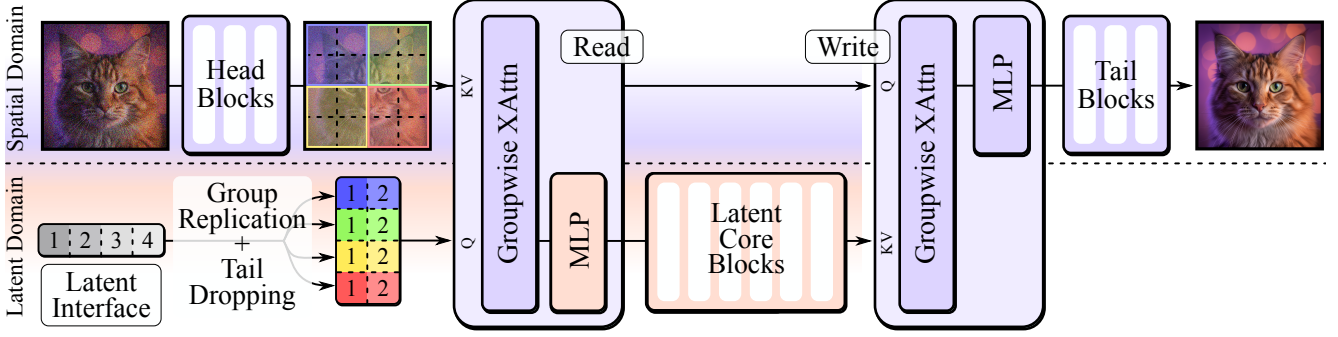


Figure 3. **Architecture of ELIT.** We extend a DiT-like generator with a variable-length set of latent tokens—the *latent interface*—and lightweight Read/Write cross-attention layers. A short spatial DiT head processes patchified inputs; Read pulls information into the latent domain where core blocks operate. Write broadcasts updated latents back to spatial tokens, and a small spatial tail produce output. Spatial tokens and latents are partitioned into corresponding groups, with cross-attention operate only within groups. During training, we randomly drop tail latents, yielding an importance-ordered interface. At inference, the number of latents serves as a user-controlled compute knob.

shown in Figure 2, DiT-B/2-Synth matches DiT-B/2 in FID, indicating no benefit from the added compute.

Attention maps in Figure 2 reveal that in DiT-B/2-Synth, zeroed tokens primarily attend to each other instead of important image regions, wasting compute. We conclude that DiT cannot reallocate computation from easy to hard regions. Such inefficiency likely also arises in natural images, where spatial regions vary in difficulty (lower or higher losses) and would benefit from compute redistribution.

3.3. Elastic Latent Interface Transformer

From spatial tokens to a variable latent interface. To allow flexible distribution of computation in DiTs, we introduce a minimal change that eliminates the fixed mapping between tokens and image patches, as shown in Figure 3. We create a latent domain by instantiating a *latent interface* of K tokens. To map the original spatial domain to the new latent domain, we use a lightweight *Read* cross-attention layer, following [12], which enables the model to select the number of latent tokens adaptively for each spatial region of $\mathbf{X}_t$, based on their difficulty. This forms a compact latent domain on which most transformer blocks operate. Finally, a lightweight *Write* cross-attention layer maps the latent updates back to the spatial grid, allowing the model to write predictions back to their locations and retain input details.

Architecture. Earlier work has shown that early and late transformer blocks in diffusion models exhibit different specializations compared to the intermediate blocks [10, 11, 55]. Therefore, we split transformer blocks into three segments:

1. **Short spatial head** (B_{in} blocks). Processes input tokens $s \in \mathbb{R}^{N \times d}$ to produce a refined spatial representation which is transferred to the latent interface. This avoids reading from raw noisy patches.
2. **Latent core** (B_{core} blocks). Variable-length latent sequence $l \in \mathbb{R}^{K \times d}$ drives most computation. We insert a Read layer $\mathcal{R}$ that pulls information from spatial to-

kens into l , then process l with standard transformer blocks in the latent domain, and finally insert a Write layer $\mathcal{W}$ that broadcasts updated latent information back to the spatial domain.

3. **Short spatial tail** (B_{out} blocks). A few blocks complete processing the written features and project them to the output velocity. This head restores fine spatial detail, noise information, and aligns features to the prediction space of $\hat{v}$.

Read and Write layers. Let $s \in \mathbb{R}^{N \times d}$ be the current spatial tokens after the spatial head and $l \in \mathbb{R}^{K \times d}$ a learnable set of initial latent tokens. The Read layer updates the latent interface via cross-attention from spatial tokens, producing output latents $l_O \in \mathbb{R}^{K \times d}$ as follows:

$$\begin{aligned} l_{CA} &= l + \text{CA}(\text{Queries} = l; \text{Keys, Values} = s), \\ l_O &= l_{CA} + \text{MLP}(l_{CA}). \end{aligned} \quad (2)$$

Conversely, the Write layer updates the spatial representation with the results of latent computations, producing updated spatial tokens $s_O \in \mathbb{R}^{N \times d}$ and is fully symmetric. We adopt pre-norm, and use adaLN-Zero [45] modulation for Read to keep the interface timestep-aware. To improve stability, we apply QK normalization inside cross-attention operations. No hidden dimensionality expansion is applied to the MLP blocks to reduce the computational overhead.

Grouped cross-attention. To reduce the cost of Read and Write operations, we partition spatial tokens into G non-overlapping groups (e.g., a regular 2D/3D grid for images/videos), and latents are partitioned accordingly in groups of $J = K/G$ latent tokens each. Latents are initialized from a set of learnable positional embeddings, which is reused across groups and encodes positional information *within* each group. This removes any dependency on a fixed input resolution: increasing spatial resolution modifies G and N , but not the number of learnable latents J . Cross attention operations attend *within* corresponding groups only. This turns the cross-attention cost from

Model	Params	ImageNet 256×256								ImageNet 512×512							
		TF @256	FID _{50K} ↓		FDD _{50K} ↓		IS ↑			TF @512	FID _{50K} ↓		FDD _{50K} ↓		IS ↑		
			−G	+G	−G	+G	−G	+G			−G	+G	−G	+G	−G	+G	
DiT-XL	675M	182	13.0	5.7	346.3	232.9	66.2	115.3		806	18.8	9.5	339.2	233.6	53.0	86.4	
⊥ ELIT	698M	188	8.2	3.8	200.2	124.5	93.0	160.1		831	11.1	4.9	175.6	106.1	80.0	134.1	
⊥ MB	698M	190	7.8 ^{40%}	3.8 ^{33%}	203.7 ^{41%}	128.6 ^{45%}	99.0 ^{50%}	167.6 ^{45%}		804	10.1 ^{46%}	4.5 ^{53%}	164.1 ^{52%}	98.2 ^{58%}	88.8 ^{68%}	147.0 ^{70%}	
UViT-XL	707M	196	8.3	<u>3.8</u>	220.2	138.0	84.4	145.1		861	11.6	5.3	202.7	125.9	72.5	117.2	
⊥ ELIT	730M	202	<u>7.5</u>	<u>3.8</u>	<u>203.8</u>	130.0	<u>95.2</u>	<u>159.7</u>		886	8.9	<u>4.2</u>	<u>155.3</u>	94.9	85.8	<u>141.0</u>	
⊥ MB	730M	204	7.1 ^{14%}	3.7 ^{3%}	203.2 ^{8%}	130.6 ^{5%}	100.3 ^{19%}	168.2 ^{16%}		858	7.7 ^{34%}	3.8 ^{28%}	135.8 ^{33%}	83.1 ^{34%}	98.0 ^{35%}	159.3 ^{36%}	
HDiT-XL	1.4B	182	12.8	5.6	361.6	247.0	68.7	119.7		776	13.0	6.0	260.3	170.5	69.4	114.2	
⊥ ELIT	1.4B	188	9.4	4.6	<u>272.2</u>	184.4	<u>89.5</u>	<u>150.2</u>		801	<u>10.1</u>	4.6	164.1	<u>112.0</u>	88.8	<u>141.0</u>	
⊥ MB	1.4B	191	9.4 ^{27%}	4.6 ^{18%}	271.8 ^{25%}	185.0 ^{25%}	92.3 ^{34%}	155.7 ^{30%}		791	9.6 ^{26%}	4.6 ^{23%}	171.2 ^{34%}	106.8 ^{37%}	94.7 ^{36%}	154.6 ^{35%}	

Table 1. **Comparative performance on ImageNet-1K at 256px and 512px resolutions.** We evaluate FID↓, FDD↓, and IS↑ without (−G) and with 0.25 CFG (+G). TFLOPs (TF) indicate single training iteration TFLOPs. Superscripts show percentage of improvement of ELIT MultiBudget (MB) relative to the baseline.

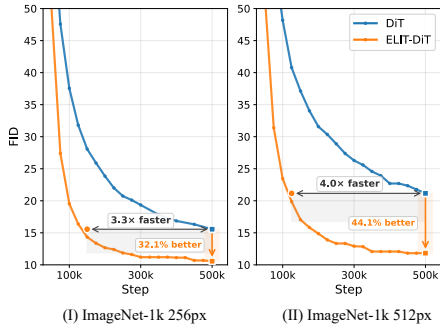


Figure 4. **Training convergence.** ELIT-DiT significantly accelerates convergence, achieving 3.3× speedup on ImageNet-1K 256px and 4.0× on ImageNet-1K 512px.

$\mathcal{O}(NK)$ into $\mathcal{O}(NK/G)$ [12].

3.4. Elastic Computation with ELIT

Spatial compute redistribution. When integrated with DiT (ELIT-DiT), our architecture enables spatial compute redistribution. In the synthetic dataset experiment described in Section 3.2, ELIT-DiT-B/2-Synth repurposes the compute from zeroed regions to enhance generation in real regions, matching the quality of the baseline trained only on the original ImageNet-1k with equivalent compute (ELIT-DiT-B/1) (see Figure 2). Attention maps of the read operations averaged over all latent tokens confirm this behavior, showing that ELIT-DiT builds its latent representation by focusing on the most informative spatial regions with the highest flow-matching loss.

Multi-budget elastic latent interface. We aim to build a multi-budget model that supports variable inference budgets. Since each latent token summarizes its group information via the read operation, a subset of tokens can still predict for the entire group, enabling budget-adaptive inference. Thus, we train an importance-ordered latent interface, where earlier tokens within each group capture globally useful information and later tokens refine details, so

that any prefix of $\tilde{J} \leq J$ tokens of the group tokens yields a valid interface associated with reduced computation (see Appendix Section K). We enforce this with a simple random prefix-keeping scheme as in [48]. At training time, we randomly sample $\tilde{J} \sim \text{Uniform}\{J_{\min}, \dots, J_{\max}\}$ once per training iteration, defining the training budget for the current iteration. The same value of $\tilde{J}$ is used across all groups. In every Read/Write and latent-core block, we keep only the first $\tilde{J}$ latents of each group and drop the subsequent tail. This process creates a consistent hierarchy where head latents are seen (and trained on) more often, forcing the model into storing the most important information in earlier latents. The generator is trained end-to-end only with the standard RF loss in Equation (1).

Asymmetric compute for improved guidance. Classifier-free guidance (CFG) [22] is a cornerstone technique in diffusion model sampling. Given a conditioning signal c and guidance scale λ , CFG redefines the velocity prediction $\hat{v}_{\text{CFG}} = (\lambda + 1)\mathcal{G}(\mathbf{X}_t | c) - \lambda\mathcal{G}(\mathbf{X}_t | \emptyset)$. While improving the quality, this comes at the cost of duplicating the number of model invocations. Recently, AutoGuidance (AG) [30] was proposed to improve guidance by replacing the unconditional model with a weak version of itself. While producing consistent improvements, it relies on the availability of a weaker model version, ideally producing artifacts that are similar to the main model [9]. Weak models are separately trained or obtained through handcrafted model corruptions [9, 23, 25]. However, in our multi-budget framework, such model is natively available by varying the inference budget defined by $\tilde{J}$. Thus, we evaluate the main term with budget $\tilde{J}$ and the guidance term with a smaller budget $\tilde{J}_w \leq \tilde{J}$. We find, however, that AG degrades metrics that reward class alignment, such as Inception Score. We thus opt to drop the class condition from the guidance term, combining the power of AG and CFG, and name the resulting guidance mechanism cheap classifier-free guidance (CCFG). The guidance mechanisms are defined as:

$$\begin{aligned}\hat{v}_{AG} &= (\lambda + 1) \mathcal{G}(\mathbf{X}_t | c; \tilde{J}) - \lambda \mathcal{G}(\mathbf{X}_t | c; \tilde{J}_w), \\ \hat{v}_{CCFG} &= (\lambda + 1) \mathcal{G}(\mathbf{X}_t | c; \tilde{J}) - \lambda \mathcal{G}(\mathbf{X}_t | \emptyset; \tilde{J}_w).\end{aligned}\quad (3)$$

This results in both improved quality and a reduced cost without any retraining or handcrafted model corruptions.

4. Experiments

4.1. Experimental setup

We demonstrate ELIT’s broad applicability by evaluating it on several popular diffusion backbones: DiT [45], U-ViT [3], and HDiT [13]. To ensure a fair comparison and evaluate architectural advantages in isolation, all baselines are built with the *same base transformer blocks*, adopt the same RF framework, and have similar train compute. Furthermore, we integrate several improvements across all baselines, including RoPE [51], and QK normalization [15]. **Training details.** We perform conditional image and video generation on ImageNet-1k [14] and Kinetics-700 [6], respectively. We train on 256px and 512px resolutions for ImageNet-1k experiments and use 29 frames, at 24 fps and 256px resolution for Kinetics-700. We use the FLUX [34] autoencoder for images and the CogVideo [60] autoencoder for videos. Main experiments are based on DiT-XL/2, while ablation studies use a DiT-B/2 model. Unless noted, we use a batch size of 256, learning rate 1×10^{-4} with 10k warmup steps, gradient clipping at 1.0, Adam [31], and EMA with $\beta = 0.9999$. Image experiments are trained for 500k steps, while video experiments are trained for 200k steps.

Evaluation metrics and protocol. We follow the evaluation protocol of [20]. For images, we report FID [21], FDD (Fréchet Distance on DINOv2 [44] features), and Inception Score (IS) [50]. For video, we report FID, FDD, and FVD [52]. Statistics are computed over 50k samples for main image experiments, while 10k samples are used for all other experiments. We use an Euler sampler with 40 steps unless otherwise noted. We use FLOPs to measure the amount of computation in all experiments and show in *Appendix* the relationship between FLOPs and forward time.

4.2. Comparison to baselines

Baseline selection and details. We use DiT [45] as a base architecture for our experiments and consider two variants: U-ViT and HDiT. U-ViT [3] adds long skip connections akin to U-Net [49]. We add ELIT read/write operations while keeping the U-ViT design untouched to obtain ELIT-U-ViT. HDiT [13] reduces tokens in intermediate blocks via PixelUnshuffle/PixelShuffle. We use a single 2×2 down-/up-sampling operation at blocks 8 and 20, respectively and double the bottleneck hidden dimensionality to match vanilla DiT FLOPs at the cost of an increased parameter count. We apply ELIT read/write operations and

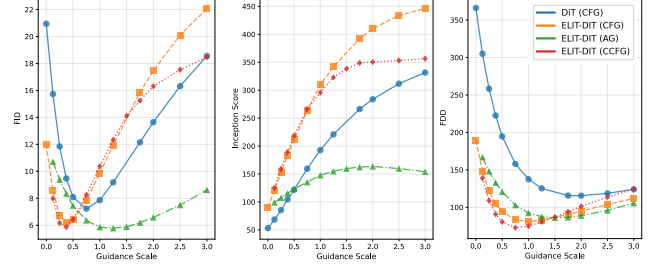


Figure 5. **Guidance strategies.** ELIT enables autoguidance out of the box by providing a well-aligned weaker model that runs at $\approx 35\%$ of the cost for the unconditional path. When paired with classifier-free guidance (CFG), denoted as cheap CFG (CCFG), it reduces overall generation cost by $\approx 33\%$ while improving quality. Compared to DiT, ELIT-DiT achieves a 19% better best FID.

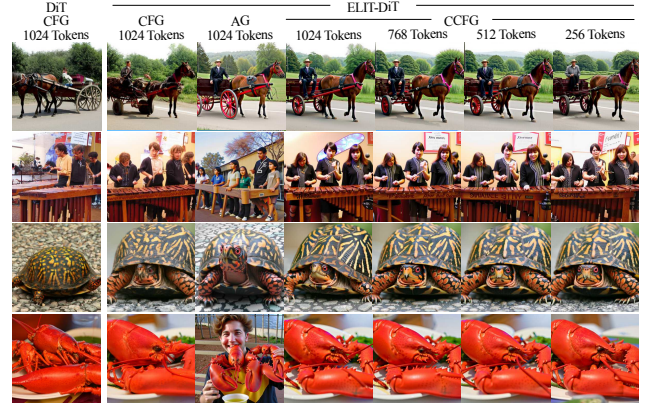


Figure 6. **Qualitative assessment of ELIT on ImageNet-1K 512px.** We compare DiT against ELIT-DiT and ablate over different guidance settings and number of latent tokens. ELIT improves structural details while allowing per-step selection of inference budget through token dropping and giving access to autoguidance (AG) and cheap classifier free guidance (CFG) for improved sampling quality and cost. See results in *Appendix Section L*.

apply group-wise downsampling in the latent space to obtain ELIT-HDiT. More details are in *Appendix Section B*.

Image generation. To disentangle the effect of dynamic compute redistribution from multi-budget training, we evaluate two variants for each baseline: ELIT, trained with a single budget matching the baselines configurations, and ELIT-MB, trained in a multi-budget setup following the tail-token dropping strategy from Section 3.4. This yields 60 budgets at 512px and 16 budgets at 256px. To account for the compute saved at shorter token lengths, we increase the batch size to 384, keeping training FLOPs comparable. We report per-iteration FLOPs for all baselines.

As shown in Table 1, ELIT, despite its simplicity, improves over all baselines under similar training FLOPs. In the multi-budget setting, ELIT-MB delivers further gains, achieving sizable improvements over DiT, U-ViT, and HDiT across all metrics, with FID reductions of 40%, 14%, and 27%, respectively. We attribute these extra gains to the

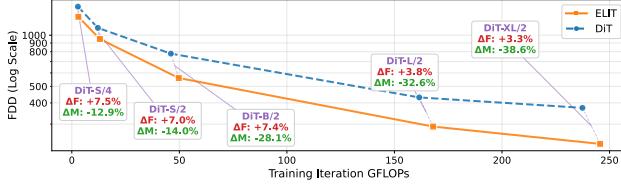


Figure 7. ELIT improves over DiT across model sizes.

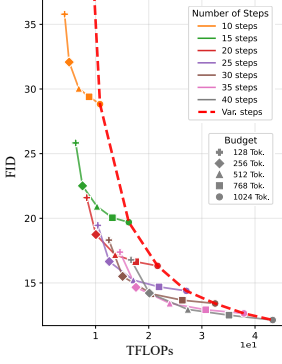


Figure 8. **Compute-quality tradeoff.** Varying inference budget in ELIT-DiT give better quality-compute tradeoff than reducing sampling steps.

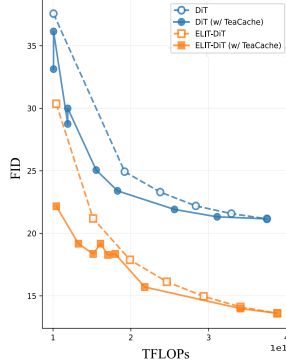


Figure 9. **TeaCache.** ELIT benefits from TeaCache similarly to DiT, yielding comparable improvements at different inference FLOPs.

importance ordering, which leads to better token semantics while benefiting from the larger effective batch size. The improvements are even more pronounced at 512px, where FID decreases by 53%, 28%, and 23% for DiT, U-ViT, and HDiT, respectively, suggesting that our method scales favorably with higher resolution, where pixel redundancy is greater and dynamic compute redistribution is more beneficial. We report in Figure 4 the convergence speed of ELIT-DiT relative to DiT at both resolutions, showing faster convergence. Figure 5 compares performance across CFG values and confirms the advantage of our method under the optimal CFG scale. Finally, Figure 6 shows qualitatively the improvements of our method over DiT. More qualitative examples are provided in Appendix Section L.

Video generation. We validate the performance of ELIT in class-conditional video generation and report the results in Table 2 where we apply grouping of spatial and latent tokens in the *frame and temporal dimension*. ELIT-DiT shows favorable results over DiT across all metrics.

Scaling across model sizes. We evaluate ELIT across model sizes from DiT-S/4 to DiT-XL/2 in Figure 7. ELIT outperforms DiT at every scale. Gains are larger for bigger models, while the relative overhead decreases, suggesting that ELIT is particularly well-suited for large-scale models.

4.3. Elastic Inference Capabilities

We analyze the ability of the model to perform inference at varied budgets using the number of retained latent tokens

Baseline	FDD _{10K} ↓		FID _{10K} ↓		FVD _{10K} ↓	
	−G	+G	−G	+G	−G	+G
DiT-XL	371.5	309.1	14.0	11.3	135.9	100.5
⊥ ELIT	277.4	222.0	13.3	10.7	116.5	90.5

Table 2. **Comparative performance on Kinetics-700 256px.** We report metrics without (−G) and with 0.25 CFG (+G).

Model	Tokens	FLOPs	Entity	Relation	Attribute	Other	Global	Avg.
Qwen-Image		1×	90.51	92.21	91.03	91.34	91.70	91.27
ELIT-Qwen-Image:								
⊥ 100% Tokens	4096	0.688×	90.30	92.18	88.97	90.34	89.18	90.45
⊥ 50% Tokens	2048	0.494×	90.15	89.94	89.05	90.09	89.06	89.81
⊥ 25% Tokens	1024	0.409×	89.31	91.87	89.71	88.28	84.79	89.79
⊥ 12.5% Tokens	512	0.369×	91.20	90.35	88.77	89.94	79.84	88.02

Table 3. Qwen-Image evaluation on DPG-Bench. Results are reported with different inference budgets.

per group $\tilde{J}$ after dropping as a knob to control the budget. **Sampling steps trade-off.** We compare our approach for controlling inference compute against naively lowering the sampling steps. As shown in Figure 8, our method delivers a superior compute-quality trade-off to varying the step count. Notably, for each FLOP target, the optimal combination of number of steps and tokens count varies, underscoring the value of models that support a continuum of inference budgets. We also demonstrate compatibility with TeaCache [39] in Figure 9, where our method attains gains comparable to the baseline when TeaCache is applied. **Efficient model guidance.** In Figure 5, we compare classifier-free guidance (CFG) with autoguidance (AG) and cheap classifier-free guidance (CCFG). Figure 6 and Appendix Section L, show qualitative examples of such guidance strategies. AG achieves comparable performance to CFG while using $\approx 33\%$ fewer FLOPs. Combining AG with CFG by dropping the class condition in the guidance term (*i.e.* CCFG) gives the best results in all metrics and delivers the same $\approx 33\%$ inference speedup.

4.4. Large Scale Multi-Budget Model

We evaluate the applicability of ELIT to large-scale generative models by applying it on top of Qwen-Image [56], which is based on a 20B MM-DiT backbone. We insert the Read and Write layers respectively after block 8 and 52. Due to the asymmetric nature of MM-DiTs and small number of text tokens (≈ 300 on average) versus 4096 image tokens at 1024px, we apply ELIT to the large image tokens stream only. Rather than outperforming the original model, a task which would require access to large-scale curated image datasets and post-training procedures matching the original ones [56], the experiment aims to demonstrate that ELIT enables stable training and multi-budget inference for large-scale MM-DiT at high resolution. Therefore, we fine-tune from Qwen-Image in a distillation setting. Specifically, we fine-tune for 60k steps at 512px resolution, using a com-

(a) Group Size Ablation					(b) Blocks Allocation Ablation				(c) Variable Budget Strategy			
Group Size	Groups	FID _{10K} ↓	FDD _{10K} ↓	IS↑	Block Alloc.	FID _{10K} ↓	FDD _{10K} ↓	IS↑	Model	FID _{10K} ↓	FDD _{10K} ↓	IS↑
ImageNet-1K 256px (16×16 tokens)					DiT-B/2				DiT	39.0	779.3	29.2
1×1	256	29.94	638.8	38.39	0-12-0	33.84	706.4	34.49	⊥ DiT + var. patch size	57.36	991.2	20.34
2×2	64	25.48	<u>546.8</u>	<u>45.66</u>	1-10-1	28.55	557.5	41.82	⊥ 25% Tokens	85.25	1181.9	13.06
4×4	16	<u>26.53</u>	531.8	45.95	2-8-2	26.53	<u>531.8</u>	45.95	ELIT-DiT + rand. drop	<u>27.0</u>	<u>540.1</u>	<u>46.3</u>
8×8	4	27.73	552.5	44.64	3-6-3	<u>25.37</u>	531.0	46.19	⊥ 25% Tokens	<u>38.6</u>	<u>718.0</u>	<u>34.5</u>
16×16	1	30.03	599.1	43.44	4-4-4	25.34	560.1	46.40	ELIT-DiT + tail drop	26.6	536.8	47.2
					5-2-5	26.95	612.1	43.15	⊥ 25% Tokens	36.3	682.1	36.4
ImageNet-1K 512px (32×32 tokens)					DiT-XL/2				(d) Batching Strategy			
1×1	1024	41.67	701.6	27.23	0-28-0	13.53	333.2	76.07	Model	FID _{10K} ↓	FDD _{10K} ↓	IS↑
2×2	256	34.50	604.0	33.99	2-24-2	12.33	239.9	86.87	(i) variable batch size	26.15	<u>537.07</u>	48.45
4×4	64	<u>31.60</u>	<u>540.1</u>	37.85	4-20-4	11.14	229.6	93.20	(ii) constant batch size	<u>26.65</u>	536.83	<u>47.18</u>
8×8	16	30.86	524.6	39.48	6-16-6	10.84	234.8	93.59				
16×16	4	31.93	545.7	<u>38.24</u>	8-12-8	10.44	237.3	95.16				
					10-8-10	<u>10.80</u>	250.1	90.11				

Table 4. **Ablations.** (a) Read/Write group size. (b) Blocks allocation to head-latent core-tail blocks. (c) Strategies for achieving variable budget inference. (d) Batching strategy for multibudget training.

bination of RF loss and a distillation loss scaled to a similar magnitude. We then fine-tune for an additional 60k steps at 1024px. We train on real images and synthetic ones generated from FLUX.1-Schnell [34] and SDXL [47].

We perform inference using the Euler sampler with 40 steps and CFG of 5.0 for the original method, while we use the faster CCFG with same weight for ELIT-Qwen-Image. Additional details are reported in *Appendix Section L*. Figure 1 and *Appendix Figure 19* show qualitative results, where ELIT-Qwen-Image cuts sampling FLOPs by up to 63%, achieving $\approx 2.7\times$ speedup while gracefully trading off speed for quality. On DPG-Bench [24], it maintains strong performance across different inference budgets, with the average score ranging from 90.45 to 88.02 at the lowest budget. We include original Qwen-Image results using the same sampling parameters for completeness. With respect to the original Qwen-Image, we observe an initial score gap of 0.82 average score points in our model.

4.5. Ablations

Group sizes. The latent group size controls how flexibly the interface can attend over spatial tokens, with larger groups enabling more opportunities for non-uniform compute. As shown in Table 4 (a), dividing the image into 16 groups performs best across 256px and 512px resolutions. Groups of 1×1 force a rigid one-to-one, spatially aligned mapping, while 16×16 spans the full 256px image and underperforms. We hypothesize that using > 1 groups provides useful coarse spatial regularization, while still permitting intra-group compute redistribution.

Blocks allocation. We vary $(B_{in}, B_{core}, B_{out})$ for DiT-B/2 and DiT-XL/2 in Table 4 (b) (i.e., the head, latent core, and tail blocks count). Optimal results occur when $\approx 67\%$ and $\approx 71\%$ of blocks are in the latent core, respectively for DiT-B and DiT-XL, with the rest split between head and tail. We use 4–20–4 for main experiments on DiT-XL.

Alternative variable-budget strategies. We evaluate other approaches for variable-budget inference in Table 4 (c). Following Anagnostidis et al. [1], Liu et al. [38], we train a DiT with two patchification layers (2×2 and 4×4) sampled uniformly during training, and set the batch size to 48 to match the baseline’s training FLOPs. In our experiments, this multi-patchification setup underperforms the standard DiT. We also replace tail-dropping strategy with random token dropping and observe a consistent performance drop.

Training Strategy. At each training step, we sample a compute budget by choosing the number of latent tokens per group, yielding variable FLOPs, which is lower on average than the baseline. To match baseline compute, we compare two strategies: (i) variable batch size scaled with lower budget, (ii) a constant batch size chosen to match baseline compute in expectation. As shown in Table 4 (d), both behave similarly, so we use the simpler constant batch size option.

5. Discussion

ELIT introduces a minimal framework that improves compute allocation in DiTs via lightweight Read/Write layers, enabling flexible compute budgets. It consistently improves image and video generation quality across architectures (DiT, U-ViT, HDiT) and resolutions, while enabling efficient compute–quality trade-offs. However, large-scale, from-scratch training benefits remain unverified. Moreover, the proposed CCFG tends to saturate images faster than CFG. Future work can explore training and inference budget schedulers that allocate different budgets across sampling steps, following prior evidence that early sampling steps require less compute [1, 29].

Acknowledgments. V.O. was partially funded by a gift from Snap Research, funding from the Ken Kennedy Institute at Rice University and NSF Award #2201710.

References

- [1] Sotiris Anagnostidis, Gregor Bachmann, Yeongmin Kim, Jonas Kohler, Markos Georgopoulos, Artsiom Sanakoyeu, Yuming Du, Albert Pumarola, Ali Thabet, and Edgar Schönfeld. Flexidit: Your diffusion transformer can easily generate high-quality samples with less compute. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 2, 8
- [2] Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *International Conference on Machine Learning (ICML)*, 2025. 2, 3
- [3] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6
- [4] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4599–4603, 2023. 15
- [5] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-All: Train One Network and Specialize it for Efficient Deployment. In *International Conference on Learning Representations (ICLR)*, 2020. 2
- [6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A Short Note on the Kinetics-700 Human Action Dataset. *arXiv*, 2019. 6
- [7] Keshigeyan Chandrasegaran, Michael Poli, Daniel Y Fu, Dongjun Kim, Lea M Hadzic, Manling Li, Agrim Gupta, Stefano Massaroli, Azalia Mirhoseini, Juan Carlos Niebles, et al. Exploring diffusion transformer designs via grafting. *arXiv preprint arXiv:2506.05340*, 2025. 13
- [8] Shuning Chang, Pichao Wang, Jiasheng Tang, Fan Wang, and Yi Yang. Sparsedit: Token sparsification for efficient diffusion transformer. *arXiv preprint arXiv:2412.06028*, 2024. 15
- [9] Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. s^2 -guidance: Stochastic self guidance for training-free enhancement of diffusion models. *arXiv preprint arXiv:2508.12880*, 2025. 5
- [10] Guanjie Chen, Xinyu Zhao, Yucheng Zhou, Tianlong Chen, and Cheng Yu. Accelerating vision diffusion transformers with skip branches. *arXiv e-prints*, pages arXiv–2411, 2024. 4
- [11] Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. Delta-dit: A training-free acceleration method tailored for diffusion transformers. *arXiv preprint arXiv:2406.01125*, 2024. 4
- [12] Ting Chen and Lala Li. FIT: Far-reaching Interleaved Transformers. *arXiv:2305.12689*, 2023. 2, 4, 5
- [13] Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z. Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *International Conference on Machine Learning (ICML)*, 2024. 6
- [14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6
- [15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633. PMLR, 2024. 3, 6
- [16] Haipeng Fang, Sheng Tang, Juan Cao, Enshuo Zhang, Fan Tang, and Tong-Yee Lee. Attend to not attended: Structure-then-detail token merging for post-training dit acceleration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18083–18092, 2025. 15
- [17] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked Diffusion Transformer is a Strong Image Synthesizer. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 16
- [18] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 2
- [19] Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Alper Canberk, Kwot Sin Lee, Vicente Ordonez, and Sergey Tulyakov. Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19373–19385, 2025. 1
- [20] Moayed Haji-Ali, Willi Menapace, Ivan Skorokhodov, Arpit Sahni, Sergey Tulyakov, Vicente Ordonez, and Aliaksandr Siarohin. Improving progressive generation with decomposable flow matching. *arXiv preprint arXiv:2506.19839*, 2025. 6
- [21] 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. In *NIPS*, 2017. 6
- [22] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5
- [23] Susung Hong. Smoothed Energy Guidance: Guiding Diffusion Models with Reduced Energy Curvature of Attention. In *Neural Information Processing Systems (NeurIPS)*, 2024. 5
- [24] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 8
- [25] Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal Skip Guidance for Enhanced Video Diffusion Sampling. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 5
- [26] Allan Jabri, David J. Fleet, and Ting Chen. Scalable Adaptive Computation for Iterative Generation. In *Proceedings*

- of the 40th International Conference on Machine Learning, pages 14569–14589. PMLR, 2023. 2
- [27] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General Perception with Iterative Attention. In *International Conference on Machine Learning (ICML)*, 2021. 2
- [28] Wongi Jeong, Kyungryeol Lee, Hoigi Seo, and Se Young Chun. Upsample what matters: Region-adaptive latent sampling for accelerated diffusion transformers. *arXiv preprint arXiv:2507.08422*, 2025. 2
- [29] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 8
- [30] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a Diffusion Model with a Bad Version of Itself. In *Neural Information Processing Systems (NIPS)*, 2024. 2, 5, 16, 18
- [31] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015. 6
- [32] Toshiaki Koike-Akino and Ye Wang. Stochastic Bottleneck: Rateless Auto-Encoder for Flexible Dimensionality Reduction. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2735–2740, 2020. 3
- [33] Felix Krause, Timy Phan, Ming Gui, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Tread: Token routing for efficient architecture-agnostic diffusion training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15703–15713, 2025. 2, 16
- [34] Black Forest Labs. Flux, 2024. 6, 8, 12
- [35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 13
- [36] Yan Li, Changyao Tian, Renqiu Xia, Ning Liao, Weiwei Guo, Junchi Yan, Hongsheng Li, Jifeng Dai, Hao Li, and Xue Yang. Learning adaptive and temporally causal video tokenization in a 1d latent space. *arXiv preprint arXiv:2505.17011*, 2025. 2, 3
- [37] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3
- [38] Dongyang Liu, Shicheng Li, Yutong Liu, Zhen Li, Kai Wang, Xinyue Li, Qi Qin, Yufei Liu, Yi Xin, Zhongyu Li, et al. Lumina-video: Efficient and flexible video generation with multi-scale next-dit. *arXiv preprint arXiv:2502.06782*, 2025. 2, 8
- [39] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep Embedding Tells: It’s Time to Cache for Video Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7353–7363, 2025. 2, 7
- [40] Xingchao Liu, Chengyue Gong, and qiang liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3
- [41] Ziming Liu, Yifan Yang, Chengruidong Zhang, Yiqi Zhang, Lili Qiu, Yang You, and Yuqing Yang. Region-adaptive sampling for diffusion transformers. *arXiv preprint arXiv:2502.10389*, 2025. 2
- [42] Wenbo Lu, Shaoyi Zheng, Yuxuan Xia, and Shengjie Wang. Toma: Token merge with attention for diffusion models. *arXiv preprint arXiv:2509.10918*, 2025. 15
- [43] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, and Sergey Tulyakov. Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7038–7048, 2024. 3
- [44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. 2, 6
- [45] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 4, 6, 12
- [46] Elia Peruzzo, Adil Karjauv, Nicu Sebe, Amir Ghodrati, and Amir Habibian. Adaptor: Adaptive token reduction for video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6365–6371, 2025. 15
- [47] 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. 8
- [48] Oren Rippel, Michael Gelbart, and Ryan Adams. Learning Ordered Representations with Nested Dropout. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1746–1754, Beijing, China, 2014. PMLR. 5
- [49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 6
- [50] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 6
- [51] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. 6

- [52] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. [6](#)
- [53] Mahtab Alizadeh Vandchali, Anastasios Kyrillidis, et al. One rank at a time: Cascading error dynamics in sequential learning. *arXiv preprint arXiv:2505.22602*, 2025. [15](#)
- [54] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 3(4):6, 2025. [1](#)
- [55] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: Decoupled Diffusion Transformer. *arXiv preprint arXiv:2504.05741*, 2025. [4](#)
- [56] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. [1](#), [7](#), [12](#), [24](#)
- [57] Haoyu Wu, Jingyi Xu, Hieu Le, and Dimitris Samaras. Importance-based token merging for efficient image and video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4983–4995, 2025. [15](#)
- [58] Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. ElasticTok: Adaptive Tokenization for Image and Video. In *International Conference on Learning Representations (ICLR)*, 2025. [3](#)
- [59] Shuo Yang, Haocheng Xi, Yilong Zhao, Muiyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *arXiv preprint arXiv:2505.18875*, 2025. [2](#)
- [60] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. [6](#)
- [61] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *International Conference on Learning Representations (ICLR)*, 2020. [3](#)
- [62] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2019. [2](#)
- [63] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An Image is Worth 32 Tokens for Reconstruction and Generation. In *Advances in Neural Information Processing Systems*, pages 128940–128966. Curran Associates, Inc., 2024. [3](#)
- [64] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Hao Luo, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dydit++: Dynamic diffusion transformers for efficient visual generation. *arXiv preprint arXiv:2504.06803*, 2025. [2](#)
- [65] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast Training of Diffusion Models with Masked Transformers. *Transactions on Machine Learning Research (TMLR)*, 2024. [2](#), [16](#)