VIT-5: VISION TRANSFORMERS FOR THE MID-2020S

Anonymous authors
Paper under double-blind review

ABSTRACT

This work presents a systematic investigation into modernizing Vision Transformer backbones by leveraging architectural advancements from the past five years. While preserving the canonical Attention–FFN structure, we conduct a component-wise refinement involving normalization, activation functions, positional encoding, gating mechanisms, and learnable tokens. These updates form a new generation of Vision Transformers, which we call ViT-5. Extensive experiments demonstrate that ViT-5 consistently outperforms state-of-the-art plain Vision Transformers across both understanding and generation benchmarks. On ImageNet-1k classification, ViT-5-Base reaches 84.2% top-1 accuracy under comparable compute, exceeding DeiT-III-Base at 83.8%. ViT-5 also serves as a stronger backbone for generative modeling: when plugged into an SiT diffusion framework, it achieves 1.84 FID versus 2.06 with a vanilla ViT backbone. Beyond headline metrics, ViT-5 exhibits improved representation learning and favorable spatial reasoning behavior, and transfers reliably across tasks. With a design aligned with contemporary foundation-model practices, ViT-5 offers a simple drop-in upgrade over vanilla ViT for mid-2020s vision backbones.

1 INTRODUCTION

Since its introduction at the end of 2020, the Vision Transformer (Dosovitskiy et al., 2021) (ViT) has substantially reshaped visual encoding paradigms. Its close architectural alignment with language transformers (Vaswani et al., 2017) has given rise to a wide range of successful vision–language systems for multimodal understanding and generation (Radford et al., 2021; Yuan et al., 2021; Yu et al., 2022a; Liu et al., 2023; Bai et al., 2025; Deng et al., 2025; Wu et al., 2025; Hurst et al., 2024). Over roughly the past five years, language models have undergone a series of systematic refinements beyond the original Transformer design (Devlin et al., 2018; Radford et al., 2018; Touvron et al., 2023a; Bai et al., 2023; Albert et al., 2023; Almazrouei et al., 2023; Team et al., 2024a; Gunasekar et al., 2023; Liu et al., 2024; Agarwal et al., 2025). These include the adoption of more advanced activation functions and normalization layers, a transition from absolute to relative positional encodings, as well as improved attention normalization and gating mechanisms. Collectively, these structural evolutions have led to substantial gains in representational capacity and training stability, and have been a major driving force behind the rapid progress of modern foundation models.

Nonetheless, in contrast to the rapid and continuous refinement of language model architectures, the core design of ViTs has remained largely unchanged since their inception. For example, the state-of-the-art plain ViT backbone DeiT-III (Touvron et al., 2022) introduces only minimal modification termed LayerScale (Touvron et al., 2021b), while otherwise preserving the original ViT architecture. Similarly, recent large-scale vision–language models such as SigLIP-2 (Tschannen et al., 2025) and Qwen3-VL (Team, 2025) continue to rely on essentially the same vanilla ViT design for visual encoding. This relative stagnation in architectural evolution raises an important question: *is the representational potential of ViTs still under-optimized?* Given the substantial gains achieved by structural refinements in language models, a natural hypothesis is whether these advancements, spanning normalization, activation functions, positional encoding, and attention mechanisms, can be systematically transferred to vision models to unlock further performance and efficiency improvements.

In this work, we present a systematic study on optimizing the architectural design of ViTs. Our investigation centers around the original plain ViT formulation, deliberately preserving its fundamental Attention–FFN backbone, while focusing on identifying the most effective variants of individual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

architectural components. Rather than proposing a radical redesign, we aim to understand how modernized design choices can be incorporated into ViTs in a principled and modular manner. We explicitly analyze and empirically validate the impact of structural refinements that have emerged since the introduction of ViT such as LayerScale (Touvron et al., 2021b), Rotary Positional Embeddings (RoPE) (Su et al., 2024), QK normalization (Team et al., 2025), and Register tokens (Darcet et al., 2024), many of which have become standard in either vision or language transformers. Through extensive experiments, we arrive at two key observations. ***First, the current ViT architecture remains under-optimized***, and refining its core components can consistently yield significant performance gains across diverse tasks. ***Second, existing architectural refinements are not strictly orthogonal***: naively combining all modern components does not necessarily lead to optimal performance, and effective modernization requires careful design.

Motivated by these findings, we introduce a next-generation Vision Transformer architecture that comprehensively refines the original design. We term this new model **ViT-5**, which denotes the incorporation of the major architectural evolutions ***over the past five years***. We present an architectural overview of ViT-5 in Figure 1. In detail, ViT-5 introduces a set of components that help stabilizing ViTs such as LayerScale, RMSNorm, and QK-Norm, as well as modules that enhance spatial reasoning such as RoPE and register tokens, while deliberately avoiding SwiGLU activations, which, despite their popularity in modern LLMs, can lead to over-gating issues (defined in Section 3.3) in vision models. Our experiments across a range of vision tasks demonstrate the strong representational capacity and good generalizability of ViT-5. On standard ImageNet-1k classification, a base-sized ViT-5 achieves 84.2% top-1 accuracy, outperforming the previous state-of-the-art plain ViT baseline, DeiT-III, which attains 83.8%. For image generation, diffusion models built on ViT-5 achieve a FID of 1.84, significantly improving upon the SiT (Ma et al., 2024) baseline's 2.06 under nearly identical computational cost. Moreover, results on dense prediction tasks, together with quantitative evaluations, further highlight ViT-5's advantages in spatial modeling and representation learning.

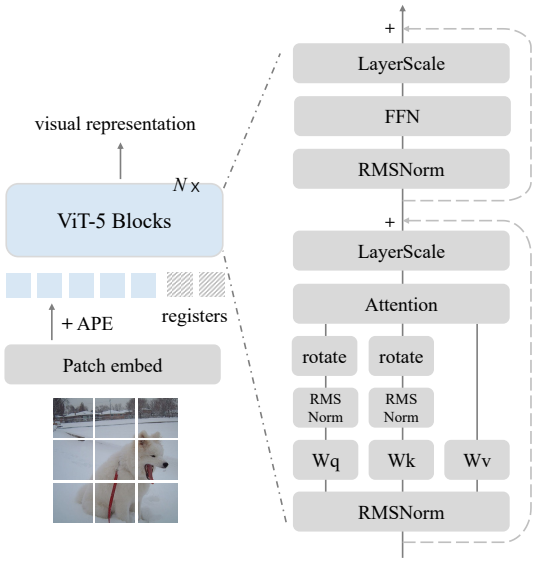


Figure 1: ***Overview of ViT-5 architecture.*** We conduct in-depth analyses and modernize the ViT architecture by refining its components including activation scaling, normalization, positional embeddings, registers, bias terms, and *etc*.

This study improves the representational capacity of ViT backbones and narrows the architectural gap between vision models and modern language models. We hope ViT-5 can facilitate more efficient construction of multimodal systems and inspire the development of unified Transformer architectures that generalize seamlessly across modalities.

2 Related Work

Modern vision backbones. The introduction of Vision Transformers marked a shift in visual representation learning from convolutional backbones (Krizhevsky et al., 2012; LeCun et al., 1998; Simonyan & Zisserman, 2015; He et al., 2016; Huang et al., 2017; Tan & Le, 2019) to transformer-based architectures (Dosovitskiy et al., 2021; Touvron et al., 2021a; 2022; Heo et al., 2024), offering a wide range of advances in visual pretraining (Bao et al., 2022; He et al., 2022; Oquab et al., 2023; Siméoni et al., 2025; Chen et al., 2021), multimodal modeling (Radford et al., 2021; Jia et al., 2021; Yu et al., 2022a; Li et al., 2022; Zhai et al., 2023; Wang et al., 2024a), and image generation (Peebles & Xie, 2023; Ma et al., 2024). Following the ViT architecture, the DeiT series (Touvron et al., 2021a;b; 2022) introduce more efficient training strategies, enabling plain Vision Transformers to match the performance of contemporary CNN backbones such as ConvNeXt (Liu et al., 2022b; Woo

et al., 2023), under comparable computational budgets. Beyond plain ViTs, a large body of work has explored hierarchical Vision Transformers (Liu et al., 2021; 2022a; Wang et al., 2021; Ryali et al., 2023; Yu et al., 2022b; Wu et al., 2021; Yang et al., 2021) and hybrid ViT–CNN architectures (Dai et al., 2021; Tu et al., 2022; Graham et al., 2021), which incorporate stronger inductive biases and often achieve superior performance on moderate-scale visual recognition benchmarks. This work does not aim to maximize absolute performance on specific benchmarks such as ImageNet (Deng et al., 2009) or COCO (Caesar et al., 2018). Instead, we focus on component-wise architectural optimization of plain Vision Transformers, with the goal of preserving their strong generalizability and scalability (Dehghani et al., 2023; Wang et al., 2025) across model sizes and diverse visual tasks.

Architectural advancements for Transformers. The Transformer (Vaswani et al., 2017) architecture has undergone a series of element-wise refinements, most prominently driven by the rapid evolution of large language models (Touvron et al., 2023a;b; Grattafiori et al., 2024; Bai et al., 2023; Team et al., 2024c; Yang et al., 2025; Team et al., 2024a;b; 2025; Almazrouei et al., 2023; Institute, 2024; Liu et al., 2024; Agarwal et al., 2025; Gunasekar et al., 2023; Abdin et al., 2024; Albert et al., 2023; AI, 2025; GLM et al., 2024). For example, early LLaMA (Touvron et al., 2023a) models introduce rotary positional embeddings (Su et al., 2024) and RMS normalization, and replace the standard MLP with SwiGLU-based feed-forward networks (Shazeer, 2020). More recently, Gemma3 (Team et al., 2025) further incorporates explicit normalization for queries and keys in self-attention, while Qwen3 (Yang et al., 2025) removes biases in the QKV projections to improve training stability and efficiency. Similar trends have also emerged in vision models. CaiT (Touvron et al., 2021b) introduces LayerScale to stabilize the optimization of deep Vision Transformers, and subsequent work has shown that incorporating register tokens (Darcet et al., 2024) can effectively mitigate activation artifacts in Transformers. Importantly, these refinements are centered around the original Transformer formulation, preserving the canonical attention–FFN inference structure without altering the overall model topology. How to effectively leverage this class of component-wise architectural improvements forms the primary focus of our study.

3 Model

3.1 Activation Scaling

Among early refinements to ViTs, it is observed that introducing a learnable scaling factor on the output of each block can significantly improve training stability and performance for deep models (Touvron et al., 2021b). Formally, given an Attention or MLP block $\mathcal{F}(\cdot)$, it applies the following transformation:

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \mathcal{F}(\mathbf{x}_l) \odot \lambda, \tag{1}$$

where $\lambda \in \mathbb{R}^d$ is a learnable scaling vector which is typically initialized to a small value (*e.g.*, 10^{-4}). This mechanism is commonly referred to as LayerScale and has been used as a default component in many modern ViT architectures such as DINO v3 (Siméoni et al., 2025).

While LayerScale has not been widely adopted in language models, we observe an intrinsic connection between LayerScale and post-normalization, which has been recognized as an important technique for stabilizing the training of deep LLMs. Formally, post-RMSNorm can be rewritten as

$$\mathbf{x}_{l+1} = (\mathbf{x}_l + \mathcal{F}(\mathbf{x}_l)) \odot \lambda_p / \text{Norm} \tag{2}$$

where

$$\text{Norm} = \text{RMS}(\mathbf{x}_l + \mathcal{F}(\mathbf{x}_l)) \tag{3}$$

and $\lambda_p \in \mathbb{R}^d$ is a scaling vector. We can find that LayerScale directly controls the scale of the block output, while post-normalization implicitly scales both the block output and residual. In our experiments, LayerScale and post-normalization lead to highly similar performance improvements (see Table 1). Given that LayerScale offers greater flexibility and lower computational overhead, ***we introduce LayerScale as a default component in ViT-5***. We hope that this observation on the functional relationship between LayerScale and post-normalization may also provide useful insights for future LLM architecture design.

model	small	base	large
ViT-5 with post-normalization	82.18	84.15	84.82
ViT-5 with LayerScale	82.16	84.16	84.86

Table 1: ***Replacing LayerScale by post-norm*** yields highly similar ImageNet top-1 accuracy across different model sizes.

LayerScale	SwiGLU	acc. (%)	FID (↓)
✗	✗	83.86	15.80
✗	✓	83.94	15.48
✓	✓	83.70	16.22
✓	✗	**84.16**	**14.57**

Table 2: ***The over-gating issue emerges*** when combining LayerScale and SwiGLU MLP. We report ImageNet accuracy for ViT-B and FID for ViT-XL.

configuration	small	base	large
no register	82.04	84.02	84.61
vanilla regs	81.95	83.90	84.37
RoPE on regs, same freq. base	82.05	84.00	84.59
RoPE on regs, high freq. base	**82.16**	**84.16**	**84.86**

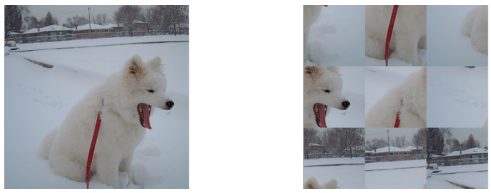
Table 3: ***Registers need relative positions.*** Vanilla registers leads to degraded performance. The issue is addressed by applying high-frequency RoPE to registers. Results denote IN1k test accuracy.

3.2 Normalization

Since influential language models such as LLaMA (Touvron et al., 2023a), PaLM (Chowdhery et al., 2023), and Gopher (Rae et al., 2021), the de facto standard in LLM architectures has largely shifted from Layer Normalization (LayerNorm) to Root Mean Square Normalization (RMSNorm). A common observation underlying this transition is that the re-scaling invariance of normalization layers dominates their practical effect. Removing the re-centering operation in LayerNorm does not degrade performance, and can even lead to slight improvements by reducing unnecessary shifting noise. We observe a similar phenomenon in ViTs. In our experiments, replacing LayerNorm with RMSNorm slightly reduces computational cost and yields a modest performance gain (*e.g.*, +0.2% top-1 accuracy on ImageNet for ViT-B). Motivated by these findings, ***we use RMSNorm throughout ViT-5***, replacing all LayerNorm layers in the original architecture.

3.3 Gated MLP

Similar to the transition from LayerNorm to RMSNorm, modern LLMs have widely utilized gated MLP architectures, in which the traditional GeLU activation is replaced by SwiGLU (Swish-Gated Linear Unit) (Shazeer, 2020). Nonetheless, in this study, we observe that combining SwiGLU MLP with LayerScale can lead to a notable performance degradation in ViTs (see Table 2). We attribute this issue to the fact that both LayerScale and gated MLPs effectively perform channel-wise filtering, which increases the sparsity of intermediate representations; when used together, their combined effect can result in excessively sparse activations. From this perspective, LayerScale can be viewed as a form of static gating, and the performance drop observed when jointly using LayerScale and SwiGLU MLP can be understood as a case of ***over-gating***. Here we draw a preliminary conclusion: at least up to the ViT-XL scale (with a hidden dimension of 1152 and 449M parameters), the combination of LayerScale and gated MLPs should be avoided. For larger-scale models with higher hidden dimensions or substantially more parameters, the over-sparsity issue may be mitigated, but we leave a systematic investigation of this regime for future work. Accordingly, ***ViT-5 uses the original MLP design with GeLU activation***.



(a) original (b) patch flipped

Figure 2: ***Undesired invariance when discarding APE***. The two images are equivalent for ViTs with only RoPE as their positional embedding. Absolute position is needed for generic vision backbones, so ViT-5 takes both APE and RoPE as default components.

3.4 Positional Encoding

Standard ViTs employ learnable absolute positional embeddings (APE), which have been shown to lack explicit relative positional modeling in complex visual reasoning tasks and to be inherently limited when handling dynamic input resolutions (Wang et al., 2024b; Agrawal et al., 2024). Following these findings, we extend rotary positional embeddings (RoPE) to the 2D setting and incorporate them into our models. Importantly, we do not discard absolute positional embeddings; instead, ***ViT-5 jointly employs both APE and 2D RoPE***. The motivation is that using relative positional encoding alone can introduce undesirable invariances. For example, under a RoPE-only formulation, patch-level flips of an image become fully invariant, as illustrated in Figure 2 , where the two input

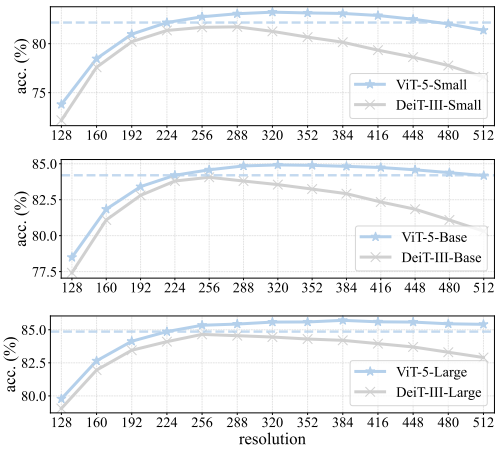


Figure 3: ***Performance at dynamic resolutions.*** All models are trained at 224^2 and then tested at different input sizes.

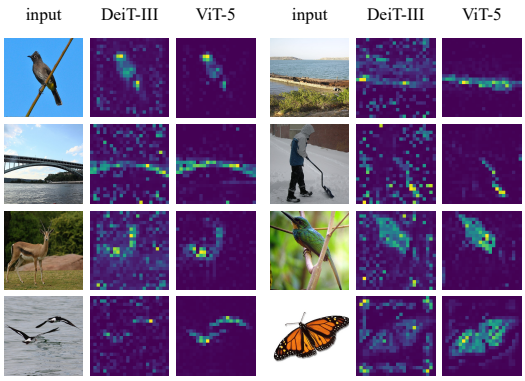


Figure 4: ***Attention visualization.*** ViT-5 exhibits improved spatial understanding with clearer self-attention activations. This improvement primarily arises from the combined effects of relative positional embeddings and registers.

images are treated as equivalent. While this behavior has limited impact on simple tasks such as image classification, it can pose potential limitations for a generic vision backbone, where absolute spatial cues may be critical for more complex visual reasoning. Figure 3 compares the ability of ViT-5 and DeiT-III (Touvron et al., 2022) to handle dynamic input resolutions. Both models are trained at a resolution of 224×224 and evaluated at different test resolutions without fine-tuning. The results show that DeiT-III, which relies solely on absolute positional embeddings, performs best only near the training resolution, and its accuracy degrades rapidly as the input size increases. In contrast, ViT-5 exhibits significantly stronger resolution robustness. Across a wide input range from 224 to 512, ViT-5 shows no noticeable performance degradation. Moreover, over a wide resolution span (*e.g.*, from 128 to 384 for ViT-5-L), increasing the input size consistently leads to performance gains, indicating improved scalability and generalization to dynamic resolutions.

3.5 Register Tokens

"Vision Transformers Need Registers" denotes a recent finding that the artifacts appear in ViTs can be effectively addressed by additional learnable tokens appended to input patch tokens (Darcet et al., 2024). In fact, the practical role of registers goes beyond their original motivation. These learnable tokens offer a flexible representation space, enabling extensions such as meta queries (Pan et al., 2025) and 1D tokenization (Yu et al., 2024) for images. To leverage these properties, ***we introduce registers as a default component in ViT-5***.

We observe that in ViTs equipped with RoPE, register tokens should also be assigned relative positional embeddings. This is because when vector rotary operations are applied only to patch tokens while leaving register tokens unrotated, the registers naturally exhibit lower cosine similarity with patches that undergo smaller rotational angles. This imbalance distorts the attention distribution and implicitly introduces an undesired positional bias. To address this issue, ***we equip register tokens with a separate 2D RoPE***, whose frequency base is significantly higher than that used for patch tokens. This design induces distinct rotational behaviors for registers and patch tokens across different channel dimensions, effectively decoupling their positional correlations and eliminating the undesired bias. Empirically, this modification stabilizes register–patch interactions and improves representation quality. Detailed performance comparison is in Table 3.

We also provide a qualitative comparison of attention activations between DeiT-III and ViT-5 in Figure 4. Consistent with prior observations (Darcet et al., 2024), register tokens effectively suppress background artifacts in attention maps, enabling the class token to attend more accurately to semantically meaningful regions of the image. In ViT-5, we observe notably cleaner and more focused feature maps, which we attribute to the combined effects of register tokens and relative positional embeddings. Both components play an important role in enhancing the spatial modeling capability of Vision Transformers. Additional visualizations are provided in the Appendix.

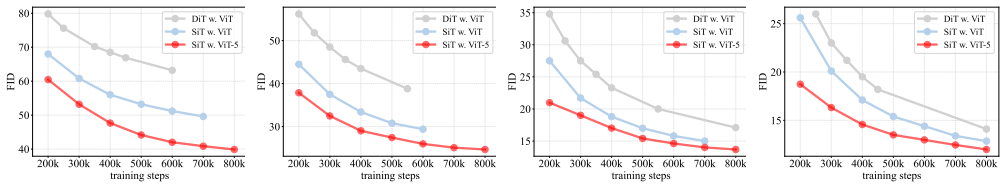


Figure 6: ***Scaling curves of FID score.*** We observe consistent performance gains and smooth scaling curves when replacing SiT's standard ViT by ViT-5. The gains are transferable across multiple model sizes (*S, B, L, and XL from left to right*) and exhibit a promising trend when training longer.

3.6 QK-Normalization

The latest LLMs such as Qwen3 (Yang et al., 2025) and Gemma3 (Team et al., 2025) have begun to reform self-attention by applying additional normalization to the query and key. Formally, this QK-Normalization mechanism has

$$Q' = \text{RMSNorm}(Q),\ K' = \text{RMSNorm}(K), \quad (4)$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(Q'K'^{T}/\sqrt{d}\right)V. \quad (5)$$

We find that QK-Normalization yields a modest performance improvement for ViTs, yet can significantly enhance training stability and reduce the occurrence of sharp loss spikes during optimization (see Figure 5). To leverage these benefits of robustness, ***we use QK-Norm as in ViT-5***.

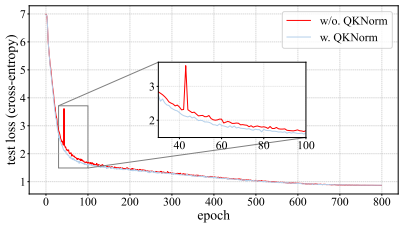


Figure 5: ***QKNorm improves stability.*** We show epoch-wise test loss of a ViT-5-Small with and without QKNorm, while the former converges smoothly and does not show spikes.

3.7 Bias Terms for QKV

We utilize bias-free RMSNorm across the entire ViT-5 models, including all pre-norm and QK-Norm layers. This highlights the perspective that self-attention relies more on weighted projections than on additive biases. Consequently, ***we remove bias terms in the QKV projection layers*** to ensure structural consistency. This modification allows QK-Norm to operate more effectively and leads to noticeable performance improvements, as detailed in Section 4.4.

4 Experiments

4.1 Image Classification

Following prior practice (Touvron et al., 2021a; 2022), we train all models from scratch on ImageNet-1k (Deng et al., 2009) by DeiT-III's recipe with small modifications. The detailed implementation is provided in the Appendix. The model configurations are summarized in Table 10. As shown in Table 4, ViT-5 consistently achieves state-of-the-art performance under comparable parameter counts and computational budgets. Notably, its performance advantage scales smoothly with both model size and input resolution. In particular, ViT-5-L attains 86.0% test accuracy with 384×384 inputs, significantly surpassing the previous state of the art for ViT (85.4%) and CNN-based models (85.5%). These results indicate that substantial and scalable gains can be achieved by systematically refining the components of ViTs. As ImageNet-1k remains the most widely used benchmark for evaluating vision backbones, the consistent improvements observed on this dataset provide strong evidence that ViT-5 offers enhanced representation learning capability and can serve as a competitive, generalized vision backbone to replace existing architectures.

4.2 Image Generation

We evaluate the transferability of ViT-5 by training it as the backbone of a Diffusion Transformer (Peebles & Xie, 2023) for image generation. We follow the same configuration as SiT, replacing the vanilla ViT backbone with ViT-5, and conduct training and evaluation on ImageNet-256 (Deng et al.,

Model	Input	#Params	FLOPs	Acc. (%)
● *ConvNeXt-T*	224^2	29M	4.5G	82.1
○ *DeiT-III-S*	224^2	22M	4.6G	81.4
● *ViT-5-S*	224^2	22M	4.7G	**82.2**
● *ConvNeXt-S*	224^2	50M	8.7G	83.1
● *ConvNeXt-B*	224^2	89M	15.4G	83.8
○ *DeiT-III-B*	224^2	87M	17.6G	83.8
● *ViT-5-B*	224^2	87M	17.9G	**84.2**
● *ConvNeXt-B*	384^2	89M	45.0G	85.1
○ *DeiT-III-B*	384^2	87M	55.5G	85.0
● *ViT-5-B*	384^2	87M	55.9G	**85.4**
● *ConvNeXt-L*	224^2	198M	34.4G	84.3
○ *DeiT-III-L*	224^2	304M	61.6G	84.5
● *ViT-5-L*	224^2	304M	62.8G	**84.9**
● *ConvNeXt-L*	384^2	198M	101.0G	85.5
○ *DeiT-III-L*	384^2	305M	191.2G	85.4
● *ViT-5-L*	384^2	305M	192.5G	**86.0**

Table 4: ***ImageNet classification results.*** We compare ViT-5 with ConvNeXt and DeiT-III, which are representative for modern CNNs and ViTs. ViT-5 attains best performance in each group of parameter count and FLOPs.

Model	#Params	Steps	FID (↓)
● *DiT with ViT-S*	33M	400k	68.4
○ *SiT with ViT-S*	33M	400k	57.6
● *SiT with ViT-5-S*	33M	400k	**47.7**
● *DiT with ViT-B*	130M	400k	43.5
○ *SiT with ViT-B*	130M	400k	33.0
● *SiT with ViT-5-B*	130M	400k	**29.0**
● *DiT with ViT-L*	458M	800k	17.1
○ *SiT with ViT-L*	458M	800k	14.9
● *SiT with ViT-5-L*	458M	800k	**13.7**
● *DiT with ViT-XL*	675M	800k	14.1
○ *SiT with ViT-XL*	675M	800k	12.6
● *SiT with ViT-5-XL*	675M	800k	**12.0**

Model	FID ↓	IS ↑	Pre ↑	Rec ↑
● *DiT ViT-XL*	2.27	278.24	0.83	0.57
○ *SiT ViT-XL*	2.06	277.50	0.83	0.59
● *SiT ViT-5-XL*	**1.84**	**282.73**	0.83	**0.60**

Table 5: ***Class-conditional image generation*** results on ImageNet-256. Top: standard training schedule. Bottom: long-horizon (7M-step) training of XL models with cfg=1.5.

2009). Implementation details are provided in the Appendix. As shown in Table 5 (top), under identical training configurations, simply replacing the original ViT backbone with ViT-5 leads to significant performance improvements over both DiT and SiT across multiple model sizes. In Table 5 (bottom), we further scale training to 7M steps and observe that ViT-5 consistently outperforms its ViT counterparts in terms of Fréchet Inception Distance (FID), Inception Score (IS), as well as Precision and Recall metrics. Figure 6 presents the scaling curves of diffusion models with different model sizes across varying training lengths. We observe that models using ViT-5 as the backbone consistently outperform their vanilla ViT counterparts, and exhibit smooth and stable scaling behavior as training progresses. Taken together, these results demonstrate that ViT-5 generalizes effectively across both visual understanding and image generation tasks, highlighting the broader impact of our component-wise architectural modernization. By systematically refining core Transformer components, ViT-5 serves as a strong and versatile backbone for diverse vision workloads.

4.3 Dense Prediction

We further evaluate ViT-5 on ADE20k (Zhou et al., 2019) for semantic segmentation using the UperNet (Xiao et al., 2018) framework. All models are trained at a resolution of 512×512 for 160k iterations, with backbones pretrained on ImageNet-1k for the same number of epochs. All other training settings are kept identical across methods and the technical details can be found in the Appendix. As shown in Table 6, ViT-5 consistently outperforms DeiT-III across all model scales. Specifically, ViT-5-Small, ViT-5-Base, and ViT-5-Large achieve 47.5%, 49.1%, and 52.0% mIoU, respectively, compared to 45.2%, 48.0%, and 49.3% for their DeiT-III counterparts under the same parameter budgets. Notably, the performance gap widens with

Backbone	Segent. head	#Parameters	mIoU (%)
○ *DeiT-III-Small*	UperNet	42M	45.2
○ *DeiT-III-Base*	UperNet	128M	48.0
○ *DeiT-III-Large*	UperNet	354M	49.3
● *ViT-5-Small*	UperNet	42M	47.5
● *ViT-5-Base*	UperNet	128M	49.1
● *ViT-5-Large*	UperNet	354M	52.0

Table 6: ***ADE20k semantic segmentation results.*** We train the models at 512×512 resolution for 160k iterations. Backbones are pretrained on ImageNet-1k for the same number of epochs. As shown, ViT-5 models consistently attain the best performance.

Configuration source	L. Scale	RMSN.	Swi.	RoPE	Reg.	QKNorm	QKV-bias	Acc.
Designing choices taken from vision transformer models:								
Vanilla (Dosovitskiy et al., 2021)	✗	✗	✗	✗	✗	✗	✓	84.85
DeiT-III (Touvron et al., 2022)	✓	✗	✗	✗	✗	✗	✓	85.39
DINO v2 (Oquab et al., 2023)	✓	✗	✗	✗	✓	✗	✓	85.42
VisionLlama (Chu et al., 2024)	✗	✗	✓	✓	✗	✗	✓	85.63
DINO v3 (Siméoni et al., 2025)	✓	✗	✗	✗	✓	✗	✗	85.50
NEPA (Xu et al., 2025)	✓	✗	✓	✓	✗	✓	✓	85.69
Designing choices taken from language transformer models:								
LLaMA (Touvron et al., 2023a)	✗	✓	✓	✓	✗	✗	✗	85.73
Qwen (Bai et al., 2023)	✗	✓	✓	✓	✗	✗	✓	85.66
GPT-oss* (Agarwal et al., 2025)	✓	✓	✓	✓	✗	✗	✗	85.64
Gemma3 (Team et al., 2025)	✗	✓	✓	✓	✗	✓	✗	85.75
ViT-5	✓	✓	✗	✓	✓	✓	✗	**86.00**

Table 7: ***Existing design choices are under-optimized for ViTs.*** We compare ViT-5 against the setup from prevalent vision models such as DeiT-III and DINO v2/v3. We also try transferring the configurations adapted from language models such as the LLaMA, Qwen, and Gemma series into ViTs (with RoPE extended to 2D). However, all these cases exhibit a clear performance gap compared with ViT-5. * indicates models that use post-norm. instead of explicit LayerScale. Acc. is ImageNet-1k top-1 accuracy for ViT-Large at 384 resolution.

model scale, indicating that the benefits of component-wise modernization become more pronounced in larger models. These results also demonstrate that ViT-5 transfers effectively to dense prediction and delivers consistent gains without task-specific architectural changes, further supporting its role as a strong and general-purpose vision backbone.

4.4 Ablation Study

In this work, we examine seven architectural components in ViTs and their corresponding design choices. As exhaustively enumerating all possible combinations would be prohibitively expensive, we focus on two complementary ablation settings. First, we compare ViT-5 with prevalent Transformer configurations that are commonly adopted in existing vision models or language models, in order to assess whether the design choices explored and widely used in prior work are optimal for vision tasks. Second, starting from the complete ViT-5 model, we individually ablate each of the seven components to isolate and quantify their respective contributions to overall performance.

Comparison to existing design. Table 7 compares ViT-5 with architectural configurations derived from prevalent Vision Transformers and modern language models. To fully assess model capacity at larger scales and higher resolutions, this ablation is conducted using ViT-Large at a resolution of 384×384. As summarized, among vision-oriented designs, models such as DeiT-III, DINO v2/v3, VisionLLaMA incorporate subsets of modern Transformer components, yet none of them simultaneously adopt all effective refinements identified in this work. As a result, their performance remains consistently below that of ViT-5, with gaps ranging from 0.31% to over 1.15% top-1 accuracy on ImageNet-1k.

We also observe that directly transferring configurations from language models to vision tasks is insufficient. Although LLaMA-, Qwen-, and Gemma-style setups introduce advanced normalization, gated MLPs, and relative positional encoding, they still underperform ViT-5 when adapted to ViTs (with RoPE extended to 2D). This highlights that architectural choices optimized for language models do not trivially translate to optimal vision performance. In contrast, ViT-5 achieves the best accuracy by systematically integrating these components with vision-specific considerations, demonstrating that existing design choices are under-optimized for Vision Transformers and motivating a principled, component-wise modernization.

Impact of individual components. Table 8 reports the results of single-component ablations on ViT-5 across different model sizes. Overall, removing any individual component leads to a consistent

Init. value	Acc.
1e-6	84.01
1e-5	84.05
1e-4	**84.16**

(a) LayerScale

Freq. base	Acc.
1e-5	84.07
1e-4	**84.16**
1e-3	84.10

(b) RoPE

#Reg.	Acc.
64	84.12
16	84.12
4	**84.16**

(c) Registers

Table 9: ***Ablation of architectural hyperparameters.*** We vary the LayerScale initialization value, the frequency base of 2D RoPE, and the number of learnable registers, and report ImageNet-1k top-1 accuracy (%) of ViT-5-B. Default setups are highlighted.

drop in accuracy compared to the complete ViT-5 configuration, confirming that each design choice contributes positively to the final performance. While the magnitude of degradation varies across components and model sizes, no single modification dominates across all settings, underscoring the complementary nature of the adopted architectural refinements.

We also observe the impact of individual architectural components is not uniform across model scales. Certain components exhibit a stronger influence on smaller models, while others become increasingly critical as model size grows. For example, replacing GeLU with SwiGLU without proper stabilization leads to a pronounced performance drop in the Small model, reflecting the higher sensitivity of compact models to gating-induced sparsity. In contrast, components such as LayerScale and 2D RoPE show more substantial effects on larger models, where deeper networks and higher representational capacity amplify the importance of training stability and relative positional modeling. Notably, removing Registers and QK-Norm results in relatively modest degradation for Small models, but leads to consistent and larger drops for Base and Large variants, indicating that these components play a more significant role as model capacity increases.

Component details. Table 9 analyzes the sensitivity of ViT-5 to several key architectural hyperparameters. We observe that LayerScale initialization has a modest impact on performance, with too small initialization values degrades the accuracy by 0.15% for a base-sized ViT-5 model. ViT-5 by default follows the practice of most LLMs's 1D RoPE to set a frequency base to 1e-5, which here showcases the best performance. In all ViT-5 models in this work, we employ four learnable registers. We observe that the number of registers only has trivial impact on predictive performance and shows good robustness to different setups.

Configuration	Small	Base	Large	Δ
remove LayerScale	82.01	83.86	84.41	-0.29
RMSNorm → LayerNorm	82.07	84.00	84.66	-0.15
*GeLU → SwiGLU**	81.87	83.70	84.35	-0.42
remove 2D RoPE	81.82	83.95	84.70	-0.24
remove Registers	82.04	84.02	84.61	-0.17
remove QK-Norm	81.99	84.05	84.78	-0.12
keep QKV-bias	82.10	84.12	84.80	-0.06
complete ViT-5	82.16	84.16	84.86	N/A

Table 8: ***Ablation study of single-component change.*** Δ denotes accuracy difference relative to ViT-5.

5 Conclusion

In this paper, we conduct a systematic modernization of Vision Transformers and show that a careful component-wise refresh can unlock substantial headroom without redesigning the overall architecture. By updating normalization and activation choices, strengthening positional encoding for spatial reasoning, and introducing lightweight gating and learnable tokens, ViT-5 becomes a practical next-generation ViT that is easier to optimize and more robust across visual settings. Extensive experiments verify consistent improvements on image classification, generation and semantic segmentation. These gains transfer across tasks, indicating that the benefits are not benchmark-specific but reflect improved representations and spatial inductive biases. More broadly, our results suggest that the evolution of ViTs can follow the same modular, best-practice-driven trajectory as modern large language model backbones, where stability and performance come from principled component choices rather than increasingly complex macro-architectures. We hope ViT-5 provides a strong, compatible foundation for mid-2020s vision and multimodal systems, and encourages the community to treat backbone modernization as a first-class, reproducible design process.

AI USE STATEMENT

Generative AI tools were used in this work solely to polish the writing and assist with LaTeX formatting. All AI-assisted edits were reviewed and verified by the authors.

ETHICS STATEMENT

The goal of this work is to advance the state of the art in Machine Learning methodologies. Given the foundational nature of our contributions, we believe the potential societal consequences align with the general progress of the field and do not present unique risks requiring specific highlighting.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 3

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. 1, 3, 8

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 4

Mistral AI. Introducing mistral 3. https://mistral.ai/news/mistral-3, 2025. 3

Q Jiang Albert, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 1, 3

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023. 1, 3

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. 1, 3, 8

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022. 2

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 3

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 2

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 2023. 4

Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified llama backbone for vision tasks. In *ECCV*, 2024. 8

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *NeurIPS*, 2021. 3, 15

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 2, 3, 5

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023. 3

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 3, 6

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2, 8, 16

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 3

Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *ICCV*, 2021. 3, 15

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. 1, 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *ECCV*, 2024. 2

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 2

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

Technology Innovation Institute. Falcon 3: Making advanced ai accessible and available to everyone, everywhere. https://falconllm.tii.ae/falcon3/index.html, 2024. 3

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 2

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 2

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1, 3

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 1

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022a. 3

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022b. 2

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, 2024. 2, 17

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 8

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 5

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 6, 15, 16

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018. 1

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 4

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, 2023. 3

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 3, 4

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2, 3, 8

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 2, 3

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 2

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024a. 1, 3

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024b. 3

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2, 3, 6, 8

Qwen Team. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024c. 3

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 2024. 15

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021a. 2, 6

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021b. 1, 2, 3

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022. 1, 2, 5, 6, 8, 17, 18

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. 1, 3, 4, 8

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. 3

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022. 3, 15

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1, 3

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, 2024a. 2

Feng Wang, Yaodong Yu, Wei Shao, Yuyin Zhou, Alan Yuille, and Cihang Xie. Scaling laws in patchification: An image is worth 50,176 tokens and more. In *ICML*, 2025. 3

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b. 4

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 3

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, 2023. 2

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

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *ICCV*, 2021. 3

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 7

Sihan Xu, Ziqiao Ma, Wenhao Chai, Xuweiyi Chen, Weiyang Jin, Joyce Chai, Saining Xie, and Stella X Yu. Next-embedding prediction makes strong vision learners. *arXiv preprint arXiv:2512.16922*, 2025. 8

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3, 6

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 3

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. *arXiv preprint arXiv:1904.00962*, 2019. 16, 17

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022a. 1, 2

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *NeurIPS*, 2024. 5

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, 2022b. 3

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 2

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 7

APPENDIX

IMAGE GENERATION EXAMPLES

Figure 7 presents qualitative examples generated by our model. Visually, the generated images exhibit high perceptual quality, with coherent global structure, well-preserved object shapes, and fine-grained texture details. The results appear visually pleasing and natural, without obvious artifacts or spatial inconsistencies. We attribute this strong visual quality not only to the improved representational capacity of the model, but also to its enhanced spatial awareness. Better spatial modeling allows the model to capture long-range dependencies and local details more accurately, leading to more consistent layouts and sharper textures. Together, these properties enable the model to generate images with richer details and improved visual fidelity.

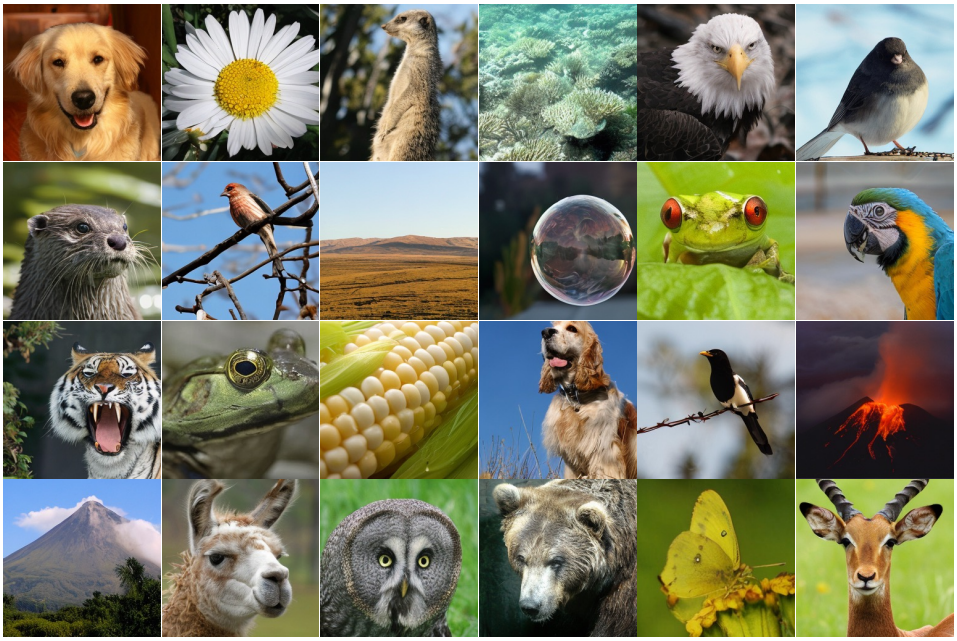


Figure 7: Image generation examples of SiT with ViT-5-XL backbone.

ARCHITECTURE DESIGN PRINCIPLES

Our design is guided by a set of high-level principles aimed at modernizing Vision Transformers through component-wise upgrades, rather than introducing new architectural paradigms. The primary objective is to improve representational capacity while preserving the strong generalizability of ViTs. The detailed principles include the following aspects: First, ***we retain the original plain ViT structure*** to ensure broad applicability across tasks and ease of integration with multimodal systems. To this end, we avoid any form of spatial downsampling within the network. Except for the initial patchification step, all intermediate feature maps maintain a fixed spatial resolution throughout the model. Second, ***we use self-attention as the only token-mixing mechanism***. While prior work has shown that incorporating convolutions can introduce stronger inductive biases and improve performance on standard benchmarks (Dai et al., 2021; Tu et al., 2022; Graham et al., 2021), we intentionally forgo these potential gains in favor of architectural simplicity and general-purpose modeling. Third, ***we do not focus on improving the patchification layer***, and instead employ the standard non-overlapping patch embedding with linear projection. This choice is motivated by the observation that many modern generation-oriented vision models no longer rely on patch embeddings for tokenization, but instead use dedicated VAEs or learned tokenizers (Peebles & Xie, 2023; Tian et al., 2024). As a result, refinements to patch embedding primarily benefit pixel-space understanding tasks and do not align with our goal of maximizing generalizability.

model	#layers	dim	#heads	#registers	#params
ViT-5-S	12	384	6	4	22M
ViT-5-B	12	768	12	4	87M
ViT-5-L	24	1024	16	4	304M
ViT-5-XL	28	1152	16	4	449M

Table 10: ***Configuration of ViT-5 models.*** The setup aligns with the practice of existing ViT models (Dosovitskiy et al., 2021; Peebles & Xie, 2023). We employ four learnable registers for all models.

More attention visualization

Figure 8 visualizes the attention maps of local tokens for DeiT-III and ViT-5 under the same experimental setting as Figure 4, except that we attend from a selected local token instead of the class token. Both models are pretrained on ImageNet-1K for the same duration and fine-tuned at a resolution of 384×384; attention visualizations are generated at the same resolution. Compared to DeiT-III, ViT-5 exhibits noticeably cleaner and more structured attention patterns. The attention of local tokens in ViT-5 is more spatially coherent and better aligned with semantically meaningful regions, while background artifacts are largely suppressed. In contrast, DeiT-III often shows more diffuse and noisy activations. Together with the class-token attention visualizations in the main text, these results qualitatively demonstrate that ViT-5 achieves a substantive improvement in spatial modeling capability, beyond numerical performance gains. The cleaner attention maps indicate more precise token interactions and more reliable spatial reasoning, highlighting the effectiveness of the component-wise modernization introduced in ViT-5.

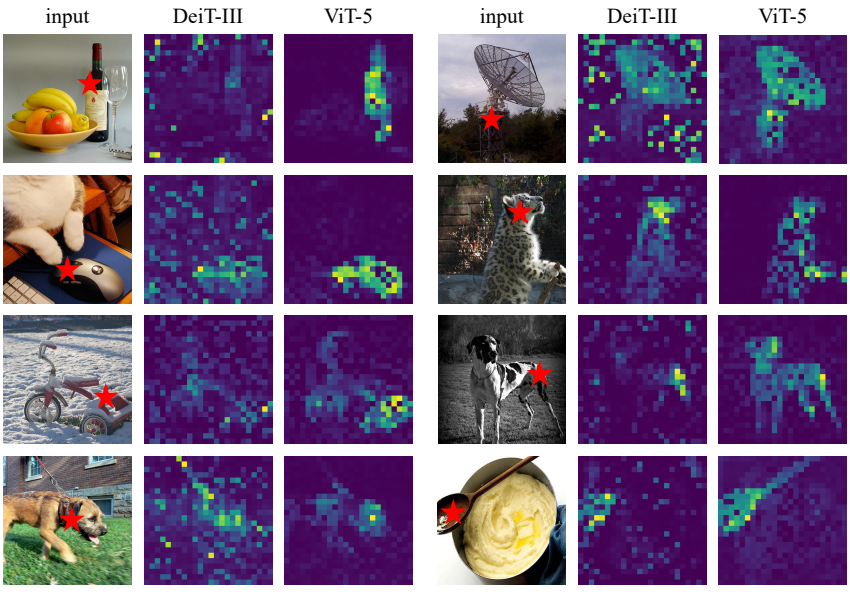


Figure 8: Attention visualization of local tokens.

Technical Details

ImageNet classification. For ImageNet-1k classification, ViT-5 follows a two-stage training pipeline, consisting of large-scale pretraining and subsequent fine-tuning. Overall, our training recipe largely follows that of DeiT-III, with several minor modifications to accommodate the architectural changes introduced in ViT-5. In the pretraining stage (Table 11), models are trained from scratch using the LAMB optimizer (You et al., 2019) with a large batch size and cosine learning rate schedule. We adopt standard data augmentations commonly used in DeiT-style training, including random resized cropping, horizontal flipping, Mixup, and CutMix, while disabling label smoothing and dropout.

Stochastic depth is applied with scale-dependent rates, and gradient clipping is enabled for training stability. In the fine-tuning stage (Table 12), we switch to the AdamW optimizer with a smaller learning rate and train for a short schedule. Fine-tuning is performed at resolutions of 224 and 384 when applicable, following common practice. Compared to pretraining, fine-tuning uses a simplified augmentation strategy with label smoothing enabled and adjusted stochastic depth rates. These settings ensure a fair and strong training baseline, while allowing the effects of ViT5's architectural refinements to be evaluated independently of aggressive recipe changes.

Image generation. For image generation, we follow the training protocol of SiT (Ma et al., 2024) and replace the vanilla ViT backbone with ViT-5. Unless otherwise specified, all hyper-parameters and optimization settings are kept identical to those in SiT to ensure a fair comparison. In particular, models are trained using the same optimizer, batch size, learning rate schedule, and data preprocessing as described in the original SiT paper. We consider two training regimes. First, we perform short training runs with a maximum of 800K steps across different model sizes to evaluate efficiency and early-stage scaling behavior. Second, to assess long-horizon scalability, we conduct a large-scale training run of ViT-5-XL for 7M steps. This long training experiment adopts a classifier-free guidance scale of 1.5 and uses the stochastic differential equation (SDE) formulation, following the SiT setup.

Semantic segmentation. For semantic segmentation, we evaluate ViT-5 on ADE20K using the UperNet framework. Our training protocol largely follows prior ViT-based segmentation works, and is kept identical across backbones to ensure fair comparison. All models are initialized from ImageNet-1k pretrained checkpoints and trained at a resolution of 512×512 for 160k iterations. We use standard data augmentation strategies for dense prediction, including random resizing, cropping, and horizontal flipping. Optimization is performed using AdamW with a cosine learning rate schedule, and stochastic depth is applied in a scale-dependent manner. No task-specific architectural modifications are introduced when replacing the vanilla ViT backbone with ViT-5.

configuration	ViT-5-small	ViT-5-base	ViT-5-large
batch size		2048	
optimizer		LAMB (You et al., 2019)	
loss		BCE	
lr	4e-3	3e-3	3e-3
lr decay		cosine	
weight decay		0.05	
warmup epochs		5	
input size	224×224	192×192	192×192
epochs	800	800	400
label smoothing ϵ		✗	
dropout		✗	
stochastic depth	0.05	0.2	0.35
repeated augmentation		✓	
gradient clipping		1.0	
horizontal flip		✓	
random resized crop		✓	
rand augment		✗	
3 augment (Touvron et al., 2022)		✓	
mixup alpha		0.8	
cutmix alpha		1.0	
random erasing prob.		✗	
color jitter		0.3	
test crop ratio		1.0	

Table 11: ViT-5 pretraining hyper-parameters on ImageNet-1k.

configuration	ViT-5-small	ViT-5-base	ViT-5-large
batch size		512	
optimizer		AdamW	
loss		CE	
lr		1e-5	
lr decay		cosine	
weight decay		0.1	
warmup epochs		5	
input size	224	224 and 384	224 and 384
epochs		20	
label smoothing ϵ		0.1	
dropout		✗	
stochastic depth	0.05	0.25	0.5
repeated augmentation		✗	
gradient clipping		✗	
horizontal flip		✓	
random resized crop		✓	
rand augment		✓	
3 augment (Touvron et al., 2022)		✗	
mixup alpha		0.8	
cutmix alpha		1.0	
random erasing prob.		0.25	
color jitter		0.3	
test crop ratio		1.0	

Table 12: ViT-5 finetuning hyper-parameters on ImageNet-1k.