Multi-Head Normalization for Wide Vision Transformers

Guoyizhe Wei[1], Feng Wang[1], Alan Yuille[1], and Rama Chellappa[1]

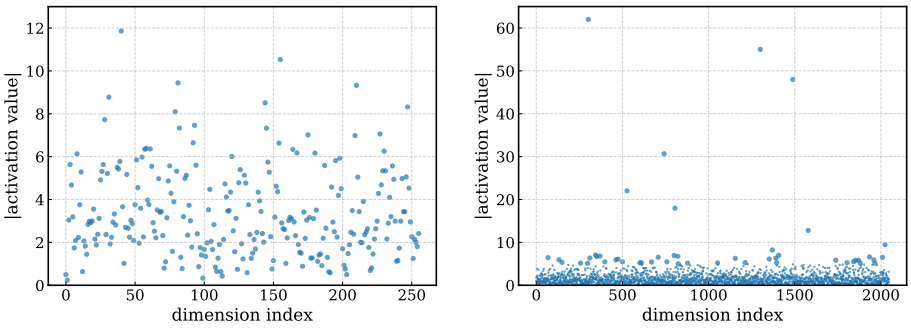
Johns Hopkins University, Baltimore, US.

Abstract. Scaling the width of Vision Transformers (ViT) often triggers severe instability issues. This work identifies that these instabilities stem from the "high-activation domination" phenomenon frequently encountered during feature normalization in high-dimensional spaces. Specifically, the norm of a high-dimensional vector can be easily dominated by a few outlier elements with disproportionately large values. Upon normalization, the remaining elements are suppressed toward zero, leading to significant gradient propagation issues. To address this, this work introduces Multi-Head RMSNorm (MH-RMSNorm), which partitions a high-dimensional feature vector into multiple chunks and applies RMSNorm to each segment independently. This simple, drop-in replacement significantly enhances the training stability and predictive performance of high-dimensional ViTs. For example, Our 1280-dimensional ViT achieves 85.5% top-1 accuracy on ImageNet-1K, whereas baseline models using standard LayerNorm or RMSNorm suffer from complete training collapse; when applied to DiT on ImageNet-256, our method achieves a 1.95 FID, significantly outperforming standard ViT counterparts.

1 Introduction

Scaling up Vision Transformers (ViTs) [4, 8, 11, 15, 18, 19, 26–28, 40] is widely regarded as a direct and effective approach to improve the model's representational capacity. In practice, parameter scaling is typically achieved through two primary axes: increasing model depth (i.e., stacking more transformer blocks) or enlarging model width (i.e., expanding the dimensionality of latent feature representations). Over the past few years, substantial progress has been made in stabilizing deep ViTs [5, 28, 44]. Techniques such as LayerScale have demonstrated that ViTs can be reliably scaled to 48 blocks or even deeper architectures without suffering from severe optimization issues [28]. In contrast, scaling the model width, namely increasing the hidden feature dimension, presents significant greater challenges. Empirically, wide ViTs frequently encounter pronounced training instability, optimization collapse, or degraded performance, even when depth scaling remains stable.

In this work, we conduct a systematic investigation into why increasing model width leads to severe optimization issues in ViTs. Our analysis reveals that the instability primarily originates from feature normalization in high-dimensional

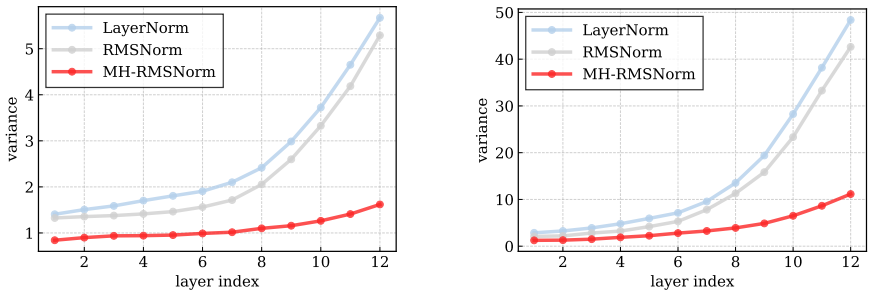


(a) 256-dimensional model. **(b)** 2048-dimensional model.

Fig. 1: ***Width-dependent activation imbalance in ViTs with LayerNorm.*** We show activation values of the final-layer CLS token for a fixed input sample. While the 256-dimensional model maintains a relatively uniform distribution, the 2048-dimensional model shows strong domination by a few dimensions, revealing increasing instability in high-dimensional normalization.

spaces. Specifically, for a high-dimensional feature vector in a ViT, we consistently observe that a small number of elements exhibit disproportionately large activations compared to the remaining dimensions. As a result, when computing LayerNorm [1] or RMSNorm [41], the overall norm of the feature vector is largely dominated by these few outlier elements. After normalization, the majority of dimensions are scaled down toward values extremely close to zero. This significantly weakens the expressive capacity of the feature representation, as most dimensions contribute negligibly after normalization. The issue becomes even more pronounced under low-precision training. In formats such as bfloat16, very small normalized values are easily truncated to zero due to limited numerical precision, causing severe gradient attenuation and, in extreme cases, complete gradient vanishing.

To provide concrete evidence of this phenomenon, we present a visual example in Figure 1. We train two 12-layer ViTs with standard LayerNorm, using hidden dimensions of 256 and 2048 respectively. For a fixed input sample, we visualize the activation values of the CLS token at the final layer across all feature dimensions. As shown in Figure 1a, the low-dimensional model (256) does not exhibit obvious outliers. The activations are distributed within a relatively balanced range, and no single dimension overwhelmingly dominates the feature magnitude. In contrast, Figure 1b shows that in the 2048-dimensional model, the feature vector is strongly dominated by a few specific elements (e.g., dimensions 302, 1300, and 1488). These elements have significantly larger activation values than the remaining dimensions. As a result, when computing LayerNorm, the overall norm of the feature vector is largely determined by these extreme activations. After dividing by the normalization factor, most dimensions are scaled down to values close to zero, severely suppressing their contribution to the representation and posting gradient vanishing issues.



(a) CLS tokens. **(b)** All tokens.

Fig. 2: ***Depth-wise amplification of activation imbalance.*** We visualize the element-wise activation statistics across transformer blocks. Both CLS tokens (global representation) and all tokens (local representation) exhibit increasing inter-dimensional variance as depth grows. The activation distribution becomes progressively more skewed in deeper layers, indicating cumulative high-activation domination.

Beyond the single example shown in Figure 1, we further observe that the high-activation outlier phenomenon accumulates with depth. Figure 2 shows that inter-dimensional variance increases rapidly in deeper layers: the CLS token, which primarily captures global information, already exhibits substantial variance growth, and this effect becomes even more severe when computed over all tokens (local representations). This suggests the imbalance is not limited to a global summary token but affects dense features throughout the network. Such accumulative high-activation behavior is consistent with previous observations in both vision and language models [6,25]. However, our analysis further highlights the underlying cause: rather than being a generic scaling artifact, the instability is primarily driven by a small number of abnormal elements that disproportionately dominate the normalization statistics.

Motivated by these observations, we propose a simple modification to the normalization procedure, which we find to be surprisingly effective. Instead of normalizing the entire feature vector as a whole, we partition it into multiple equal-sized chunks and apply normalization independently to each chunk. We refer to this mechanism as **Multi-Head Normalization**. Concretely, given a high-dimensional feature vector, we evenly divide it into n sub-vectors and perform LayerNorm or RMSNorm on each sub-vector separately. The normalized chunks are then concatenated back to form the final output representation. This design alleviates high-activation domination in two important ways. First, by splitting the feature vector, the influence of a single extreme activation is confined within its own chunk, preventing it from determining the normalization scale of the entire representation. Second, by suppressing activation imbalance at shallow layers, the subsequent accumulation effect across depth is significantly weakened. As a result, activation variance grows more slowly throughout the network, leading to improved optimization stability and more reliable gradient propagation.

Extensive experiments demonstrate the effectiveness of MH-Norm across both image understanding and generation tasks. We replace standard LayerNorm with MH-Norm in various ViT architectures and consistently observe significant improvements. For example, on ImageNet-1K classification, our method improves a ViT-H model by 0.29% top-1 accuracy. On ImageNet-256 image generation, replacing the normalization in DiT reduces the FID from 2.27 to 1.95 — these gains come at no additional computational cost, as MH-Norm introduces only a lightweight reorganization of feature normalization. More importantly, we find that MH-Norm fundamentally improves width scalability. In practice, most existing ViT models remain around 1024 hidden dimensions, as further width scaling often leads to training instability or collapse. In contrast, with MH-Norm, we are able to smoothly scale model width up to 4096 dimensions while continuing to obtain stable optimization and consistent performance gains. Overall, these results suggest that normalization plays a critical role in width scaling, and that a simple structural modification can unlock stable training for substantially wider Vision Transformers.

2 Related Work

Normalization in Transformers. Transformers [31] and their vision adaptations such as ViT [8], DeiT [26,27], Swin [15], MAE [11], DINO/DINOv2 [4,18], and SigLIP [40] have established strong baselines for visual recognition. Normalization is a key ingredient for stable optimization. LayerNorm [1] is the default in most Transformer variants, while RMSNorm [41] removes mean-centering and is widely adopted in large-scale models (e.g., T5 [22] and recent open-weight LLMs [3, 10, 21, 29, 30, 37, 38]) and has efficient implementations such as FlashNorm [9]. Normalization placement (Pre-LN vs. Post-LN) is also crucial; analyses of gradient propagation and initialization help explain why Pre-LN typically improves stability, while Post-LN often requires additional stabilization [36]. Beyond placement, several works insert extra normalization operations to improve conditioning, reduce gradient mismatch, or prevent activation/gradient blow-up, including NormFormer-style extra norms [24] and sandwich normalization for large Transformer training [7]. Alternative normalization formulations include ScaleNorm/FixNorm for self-attention normalization [17] and PowerNorm, which revisits batch-statistics-based normalization for Transformers [23].

Stabilizing Deep Vision Transformers. Training deeper ViTs can trigger collapse-like degeneration and reduced attention/representation diversity. DeepViT reports attention collapse and proposes re-attention to restore diversity [44]. CaiT enables deeper image Transformers through architectural refinements and LayerScale [28], and DiverseViT argues that reducing redundancy at multiple levels leads to stronger ViT representations [5]. Other stabilization mechanisms include stochastic depth (DropPath) [13], residual re-parameterizations such as ReZero [2], and theoretically motivated deep scaling rules such as DeepNorm in DeepNet [33]. Related efforts also explore alternative normalization/stabilization designs such as Query-Key Normalization [12], Fixup initialization [42], and more

recent proposals including LipsFormer [20] and TaperNorm [16]. Swin Transformer V2 further improves stability at scale via residual-post-norm and cosine attention [14]. Our approach is complementary: rather than modifying attention computation or introducing new residual parameterizations, we focus on the normalization statistic itself.

Group-wise and Head-wise Normalization. Group-wise normalization is a classical way to decouple statistics across channel groups. GroupNorm [34] normalizes within channel groups and is robust to batch size. Related analyses also point out that high-magnitude outliers can dominate normalization statistics in Transformers [6, 25]. In Transformers, head-wise operations appear in the form of per-head parameters or head-wise scaling [24]. We build on this intuition but target RMS-based normalization in multi-head attention: MH-RMSNorm computes RMS statistics on head-aligned groups (e.g., value-head groups), directly addressing head competition induced by global statistics and pairing naturally with a branch-end sandwich placement.

3 Method

3.1 Preliminaries

Let $X \in \mathbb{R}^{B\times T\times D}$ denote a batch of token representations, where B is the batch size, T is the number of tokens, and D is the hidden dimension. We write $x = X_{b,t} \in \mathbb{R}^D$ for a single token embedding. LayerNorm [1] normalizes each token by its mean and variance across the feature dimension:

$$\mu(x) = \frac{1}{D}\sum_{i=1}^{D} x_i, \quad \sigma^2(x) = \frac{1}{D}\sum_{i=1}^{D}\big(x_i - \mu(x)\big)^2, \tag{1}$$

$$\mathrm{LN}(x) = \frac{x - \mu(x)}{\sqrt{\sigma^2(x) + \epsilon}} \odot \gamma + \beta, \tag{2}$$

where ϵ is a small constant and $\gamma, \beta \in \mathbb{R}^D$ are learnable scale and bias.

Recently, it has been observed that the additive bias term in LayerNorm is often not critical for model expressivity and may introduce unnecessary shifts/noise in practice. As a result, many modern architectures—including recent open-weight LLM families such as LLaMA and Qwen [3,10,21,29,30,37,38], as well as contemporary vision Transformers (e.g., ViT-5 [32])—commonly replace LayerNorm with its simplified variant RMSNorm [9,41]. Formally, RMSNorm removes mean-centering and the additive bias, and instead uses the root-mean-square statistic:

$$\mathrm{RMS}(x) = \sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2 + \epsilon}, \quad \mathrm{RMSNorm}(x) = \frac{x}{\mathrm{RMS}(x)} \odot \gamma, \tag{3}$$

where $\gamma \in \mathbb{R}^D$ is a learnable scale. Using $\sigma^2(x) = \frac{1}{D}\sum_i x_i^2 - \mu(x)^2$, RMSNorm can be viewed as a simplified LayerNorm that drops mean-centering; when token

features are approximately zero-mean (i.e., $|\mu(x)| \ll \mathrm{RMS}(x)$), the two normalizers yield similar scaling behavior. In this paper, we follow this recent trend and use RMSNorm as the default choice. Unless otherwise stated, our *multi-head norm* refers to *multi-head RMSNorm*, detailed in Sec. 3.2.

3.2 Multi-Head Normalization

Given a token embedding $x \in \mathbb{R}^D$, we partition its feature dimension into H heads, where H is the number of attention heads and $D = H \cdot d$ with per-head dimension d. We write $x = [x^{(1)}, \ldots, x^{(H)}]$ with $x^{(h)} \in \mathbb{R}^d$. Multi-head RMSNorm computes an RMS statistic independently for each head,

$$\mathrm{RMS}\big(x^{(h)}\big) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}\big(x_i^{(h)}\big)^2 + \epsilon}, \tag{4}$$

and normalizes head-wise as

$$\text{MH-RMSNorm}(x)^{(h)} = \frac{x^{(h)}}{\mathrm{RMS}(x^{(h)})} \odot \gamma^{(h)}, \tag{5}$$

where $\gamma^{(h)} \in \mathbb{R}^d$ is a learnable scale for head h. Intuitively, Multi-Head Normalization normalizes each head in its own scale, preventing a few high-magnitude dimensions (or heads) from dominating the normalization statistics of the entire token representation.

3.3 Theoretical Analysis of Activation Domination

Here we provide a theoretical analysis explaining 1) why high-dimensional representations are prone to activation domination, 2) how such domination degrades representation quality and gradient propagation, and 3) why Multi-Head RMSNorm mitigates these issues. We present detailed derivations in Appendix.

Extreme Value Growth in High Dimensions. Let $x \in \mathbb{R}^D$ denote a token representation before normalization. Assume all elements in the feature vector are independent with

$$x_i \sim \mathcal{D}, \quad \mathbb{E}[x_i] = 0, \quad \mathrm{Var}(x_i) = \sigma^2. \tag{6}$$

Ignoring ϵ for clarity, RMSNorm uses

$$\mathrm{RMS}(x) = \sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}, \tag{7}$$

where by the law of large numbers, we have,

$$\mathrm{RMS}(x) \to \sigma. \tag{8}$$

According to extreme value theory, we have,

$$\mathbb{E}\big[\max_i |x_i|\big] \approx \sigma\sqrt{2\log D}. \tag{9}$$

Thus,

$$\frac{\max_i |x_i|}{\mathrm{RMS}(x)} \sim \sqrt{2\log D}, \tag{10}$$

which increases with dimensionality. Hence, high-dimensional vectors are increasingly likely to contain extreme activations that dominate the squared norm.

Lemma 1 (Representation Suppression). *Suppose a subset S with $|S| = k \ll D$ satisfies*

$$x_j^2 = M^2 \ (j \in S), \quad x_i^2 = \sigma^2 \ (i \notin S), \quad M^2 \gg \sigma^2. \tag{11}$$

Then after RMS normalization,

$$\hat{x}_i = \frac{x_i}{\mathrm{RMS}(x)}, \tag{12}$$

the energy outside S satisfies

$$\frac{\sum_{i\notin S} \hat{x}_i^2}{\sum_{i=1}^{D} \hat{x}_i^2} = \mathcal{O}\left(\frac{D\sigma^2}{kM^2}\right). \tag{13}$$

As $M \to \infty$, this ratio approaches zero.

Implication. Most representational energy concentrates in the dominant subset S, effectively reducing the intrinsic dimensionality of $\hat{x}$.

Lemma 2 (Gradient Attenuation). *For non-dominant dimensions $i \notin S$,*

$$\frac{\partial \hat{x}_i}{\partial x_i} \approx \frac{1}{\mathrm{RMS}(x)} = \mathcal{O}\left(\frac{1}{M}\right). \tag{14}$$

Hence,

$$\left|\frac{\partial \mathcal{L}}{\partial x_i}\right| = \mathcal{O}\left(\frac{1}{M}\right), \tag{15}$$

which vanishes as M grows.

Mitigation via Multi-Head RMSNorm. Partition x into H heads with dimension $d = D/H$. Normalization within each head uses

$$\mathrm{RMS}_h\big(x^{(h)}\big) = \sqrt{\frac{1}{d}\sum_{i=1}^{d} \big(x_i^{(h)}\big)^2}. \tag{16}$$

If domination is confined within individual heads, suppression scales with $\sqrt{d}$ instead of $\sqrt{D}$. Since $d \ll D$, normalization effects are localized and gradient magnitudes are correspondingly larger.

Proposition 1. *Multi-Head RMSNorm reduces global activation domination by restricting normalization statistics to lower-dimensional subspaces, thereby preserving representational diversity and improving gradient propagation.*

Model	Width	# Params	Normalization	Accuracy (%)	mIoU (%)
ViT-B	768	86M	LayerNorm	83.8	48.1
			RMSNorm	83.8	48.1
			MH-RMSNorm	**83.9**	**48.3**
ViT-L	1024	304M	LayerNorm	84.5	49.2
			RMSNorm	84.5	49.3
			MH-RMSNorm	**84.7**	**49.6**
ViT-H	1280	632M	LayerNorm	85.1	50.5
			RMSNorm	85.2	50.6
			MH-RMSNorm	**85.5**	**51.0**

Table 1: ImageNet-1K classification and ADE20K segmentation with different normalizers. We compare LayerNorm, RMSNorm, and our MH-RMSNorm under identical training settings. We report ImageNet-1K top-1 accuracy and ADE20K mIoU, and observe consistent improvements from MH-RMSNorm, with larger gains in wider models.

4 Experiments

4.1 Performance on Standard Benchmarks

We first benchmark MH-RMSNorm on standard vision benchmarks to evaluate its predictive performance. To isolate normalization as the only variable, we keep all other components unchanged and replace the original normalizer with MH-RMSNorm, training all models under identical settings.

Image classification. We follow the DeiT-III training recipe [27] for supervised ImageNet-1K classification, which uses the LAMB optimizer [39] with cosine learning-rate decay and warmup, together with standard data augmentation and mixup strategies. We train ViT-Base, ViT-Large, and ViT-Huge with hidden dimensions $D \in \{768, 1024, 1280\}$, using 800 epochs for ViT-Base and 400 epochs for ViT-Large/Huge. For MH-RMSNorm, we set the number of heads to $H \in \{3, 4, 5\}$ for Base/Large/Huge respectively, which results in the same per-head dimension $d = D/H = 256$. Table 1 shows that MH-RMSNorm consistently matches or improves upon LayerNorm and RMSNorm across model sizes. While the absolute gains on ImageNet-1K are modest (e.g., 83.8% to 83.9% on ViT-B, and 85.2% to 85.5% on ViT-H compared with RMSNorm), the improvement becomes more noticeable as width increases, which supports our hypothesis that global normalization in high dimensions is more likely to be dominated by a small set of outlier activations and that head-decoupled normalization alleviates this domination.

Semantic Segmentation. We further benchmark our method on ADE20K [43] for semantic segmentation within the UperNet [35] framework. Following a controlled protocol, all models are trained for 160k iterations at a 512×512 input resolution, using backbones pretrained on ImageNet-1K for the same number of epochs. Unless otherwise noted, we keep the remaining training hyperparameters and data processing pipeline identical across all compared methods. As shown in Table 1, MH-RMSNorm yields consistent improvements over global normalizers,

Model	Width	# Params	Steps	Normalization	FID ↓
DiT-B/2	768	130M	400k	LayerNorm	43.5
				RMSNorm	43.2
				MH-RMSNorm	**42.6**
DiT-L/2	1024	458M	800k	LayerNorm	17.1
				RMSNorm	17.0
				MH-RMSNorm	**15.9**
DiT-XL/2	1152	675M	800k	LayerNorm	14.1
				RMSNorm	13.8
				MH-RMSNorm	**13.0**

Table 2: ImageNet-256 class-conditional generation with different normalizers. We report FID (lower is better) for DiT models under identical training settings.

Normalization	FID ↓	IS ↑	Precision ↑	Recall ↑
LayerNorm	2.27	278.24	0.83	0.57
RMSNorm	2.20	279.01	0.83	0.57
MH-RMSNorm	**1.95**	**281.15**	**0.83**	**0.59**

Table 3: ImageNet-256 class-conditional generation results for DiT-XL/2 trained for 7M steps with classifier-free guidance (CFG) scale 1.5. We report FID (lower is better), Inception Score (IS, higher is better), and precision/recall (higher is better).

and the gain is more pronounced than in classification (e.g., from 50.6 to 51.0 mIoU on ViT-H compared with RMSNorm). We hypothesize this is because segmentation operates on larger and denser feature maps, where activation outliers are more likely to emerge and dominate global normalization statistics; head-decoupled normalization mitigates such domination and therefore yields larger benefits in this more challenging setting.

Image generation. Figure 3 visualizes representative samples, showing that MH-RMSNorm enables high-quality generations with clear object structure, fine-grained textures, and coherent global composition under class conditioning. Notably, samples remain sharp while maintaining diversity across instances of the same class, suggesting improved coverage without sacrificing fidelity. Quantitatively, Tables 2 and 3 show that MH-RMSNorm also benefits diffusion-based generation. In Table 2, the effect is consistent across DiT scales: for example, DiT-B/2 improves from 43.2 (RMSNorm) to 42.6 FID, while the wider DiT-XL/2 improves from 13.8 to 13.0, indicating that wider diffusion backbones are more sensitive to how normalization aggregates statistics. This is aligned with our analysis: diffusion training introduces strong stochasticity, and in high-dimensional layers a few outlier activations can disproportionately influence a global normalizer, making optimization more brittle; normalizing within heads reduces the chance that a single extreme dimension suppresses the rest. In the stronger DiT-XL/2 setting (Table 3), MH-RMSNorm improves FID from 2.20 to 1.95 and increases recall from 0.57 to 0.59 while keeping precision at 0.83, sug-

Fig. 3: ***Image generation samples on ImageNet-256.*** Qualitative samples generated by DiT-XL/2 with MH-RMSNorm (CFG scale 1.5). The results exhibit strong semantic alignment with the class condition, high visual fidelity, and diverse appearances across categories, with few common diffusion artifacts.

gesting that head-decoupled normalization primarily improves coverage of the data distribution (less mode dropping) without sacrificing sample sharpness.

4.2 Scaling to Wider Models

We conduct a dedicated width-scaling study to stress-test the stability of normalization in high-dimensional ViTs. The goal is to evaluate whether MH-RMSNorm can serve as a drop-in replacement that enables ViTs to reliably scale to substantially wider representations, where conventional global normalization methods often exhibit optimization degradation or even training collapse. Under the same model configuration and training recipe, we progressively scale the ViT width from 256 to 4096 in order to isolate the effect of width scaling on normalization stability. Specifically, we follow the standard ViT-Base architecture and fix the depth to 12 transformer blocks, keeping all architectural components unchanged while only varying the latent feature dimension (i.e., the width).

Experiment configuration. For image classification, we follow the training recipe of DeiT-III. Models are first pretrained for 400 epochs on ImageNet-1K using a resolution of 192×192, and then finetuned for 20 epochs at 224×224 resolution. For image generation, we follow the experimental setup of DiT on ImageNet-256. To improve experimental efficiency, we train all models for 200k optimization steps while keeping the remaining training configurations identical

to the original DiT implementation. Performance is evaluated using top-1 accuracy for classification and FID for generation. For MH-RMSNorm, the number of heads is determined by the feature dimension. We fix the per-head dimension (i.e., the chunk size) to 256 and partition the feature vector accordingly. For example, a model with width 1024 uses 4 heads, while a model with width 4096 uses 16 heads. This design keeps the normalization subspace size consistent across different model widths.

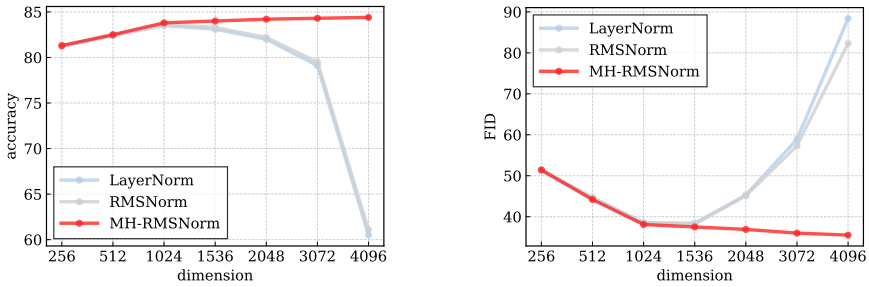
Results. The width scaling results are summarized in Figure 4. For image classification, we observe a clear divergence between conventional global normalization methods and MH-RMSNorm as the model width increases. As shown in Figure 4, both LayerNorm and RMSNorm exhibit similar scaling behavior. Their performance improves as the width increases from 256 to 1024, reaching a peak around 1024–1536 dimensions. However, further increasing the width leads to noticeable performance degradation. In particular, when the width reaches 3072 and 4096, both methods suffer from severe optimization instability, and the top-1 accuracy drops dramatically. In contrast, MH-RMSNorm demonstrates consistently stable scaling behavior. As the width increases, the model continues to improve and maintains strong performance even at extremely large dimensions. Notably, while LayerNorm and RMSNorm collapse at width 4096, MH-RMSNorm still achieves the best accuracy in this regime. This result suggests that the instability observed in wide ViTs is closely related to the global normalization mechanism, and that localizing normalization statistics via MH-RMSNorm effectively alleviates this limitation.

A similar trend can be observed in the image generation experiments. As the model width increases, both LayerNorm and RMSNorm initially improve generation quality, but their performance quickly deteriorates beyond moderate widths. The FID score starts to increase significantly after width 1536 and rapidly diverges when the width reaches 3072 and 4096, indicating severe training instability. In contrast, MH-RMSNorm exhibits remarkably stable scaling behavior across the entire width range. As the model becomes wider, the generation quality steadily improves, with FID decreasing consistently up to width 4096. This trend suggests that wider generative transformers indeed possess greater modeling capacity, but conventional global normalization prevents the model from effectively exploiting this capacity. By normalizing features within smaller subspaces, MH-RMSNorm stabilizes optimization and enables wider models to fully realize their representational potential.

4.3 Ablation Study

We perform ablations to understand which design choices are critical for MH-RMSNorm, and to verify that the gains are not merely from increasing the number of normalization "groups". Unless otherwise specified, we use the width-scaling setting in Section 4.2 (Figure 4) and report results at representative widths where global normalization begins to degrade.

Number of heads and chunk size. We ablate the number of heads by varying the chunk size while keeping all other settings fixed. Table 4 shows that



(a) ImageNet-1K classification scaling. **(b)** ImageNet-256 generation scaling.

Fig. 4: Width scaling from 256 to 4096. We scale the model width while keeping the depth fixed (12 blocks) and all other settings unchanged, and report top-1 accuracy on ImageNet-1K and FID on ImageNet-256.

Setting	Width D	H	Chunk d	Top-1 (%) ↑	Notes
RMSNorm (global)	4096	1	4096	61.1	unstable
MH-RMSNorm	4096	8	512	84.2	fewer heads
MH-RMSNorm	4096	16	256	**84.4**	default
MH-RMSNorm	4096	32	128	84.3	more heads
MH-RMSNorm	4096	64	64	83.5	too fragmented

Table 4: Ablation on the number of heads H (chunk size d) for width scaling. Results are on ImageNet-1K at width 4096 with depth fixed to 12. Our default setup is highlighted in orange. The best result is bolded.

global RMSNorm becomes unreliable at extreme width, dropping to 61.1% top-1 at width 4096. Once we partition features into a few chunks, performance largely recovers: $d = 512$ already reaches 84.2%, and our default $d = 256$ gives the best result at 84.4%. Making chunks moderately smaller keeps accuracy nearly unchanged (84.3% at $d = 128$), while overly small chunks start to hurt (83.5% at $d = 64$). This also explains why the issue is easy to miss in common ViT regimes, where widths are typically around 1k or below and global normalization can still appear well-behaved. Our ablation isolates that the failure mode emerges primarily in the ultra-wide setting, and that selecting a moderate chunk size effectively prevents outliers from dominating the normalization statistics. Overall, MH-RMSNorm is robust across a wide range, and $d = 256$ provides a good balance between localization and coupling.

Comparison to other normalization formats. We compare different normalizers under the same multi-head partition to disentangle the effect of local statistics from the choice of normalization family. Table 5 shows that global normalizers fail to scale to width 4096, yielding only about 60% top-1 and extremely poor FID. Simply introducing a multi-head partition already restores most of the performance, as MH-LayerNorm reaches 84.2% top-1 with a large FID im-

Normalization	Width D	Partition	Top-1 (%) ↑	FID ↓
LayerNorm (global)	4096	N/A	60.5	88.4
RMSNorm (global)	4096	N/A	61.1	82.3
MH-LayerNorm	4096	$d = 256$	84.2	36.1
MH-LayerNorm (shared γ, β)	4096	$d = 256$	83.9	37.5
MH-RMSNorm	4096	$d = 256$	**84.4**	**35.5**
MH-RMSNorm (shared γ)	4096	$d = 256$	84.0	35.9

Table 5: Ablation across normalization families with multi-head partition. We compare standard multi-head normalizers and variants that share learnable affine parameters across heads. Top-1 is evaluated on ImageNet-1K at width 4096 (depth 12). FID is evaluated on ImageNet-256. Our default setup is highlighted in orange.

provement. Replacing LayerNorm with RMS-based statistics further helps, and MH-RMSNorm attains the best results on both classification and generation. A notable takeaway is that the ranking is consistent across discriminative and generative metrics, suggesting that the underlying issue is not task-specific but rooted in optimization dynamics of wide transformers. Moreover, the gap between shared-parameter and head-wise-parameter variants indicates that once statistics are localized, head-wise affine parameters become a lightweight way to adapt to heterogeneous activation scales across subspaces.

Where to apply MH-RMSNorm. We first study whether MH-RMSNorm mainly helps a particular computation path by applying it to the attention blocks or the MLP blocks only. Table 6 shows that the two partial replacements behave almost identically: using MH-RMSNorm only in attention blocks achieves 84.0% top-1 and 37.3 FID, while applying it only in MLP blocks also reaches 84.0% top-1 with 37.2 FID. This symmetry suggests that the gain is not tied to a specific operator, but instead comes from mitigating activation domination in the normalization itself. Replacing both branches provides a further, consistent improvement to 84.2% top-1 and 36.9 FID.

We next ablate which transformer layers use MH-RMSNorm. As shown in Table 7, applying MH-RMSNorm only in shallow layers gives limited benefit (83.1% top-1 and 41.0 FID), while moving it to deeper layers yields stronger gains (83.9% top-1 and 38.4 FID for layers 9–12). This trend matches our analysis in Section 1 that activation outliers are weaker in early blocks but amplify with depth, making deeper normalization layers more critical. At the same time, using MH-RMSNorm in the middle of the network is already helpful (83.7% top-1 and 39.6 FID), indicating that intervening earlier can reduce the accumulation effect. Overall, applying MH-RMSNorm throughout the network performs best, reaching 84.2% top-1 and 36.9 FID.

5 Conclusion

We study why scaling Vision Transformers in width often triggers optimization instability and performance degradation, and identify a recurrent failure mode in feature normalization: in high-dimensional representations, a small number

Normalization type		Width	Top-1 (%) ↑	FID ↓	Notes
Attention blocks	MLP blocks				
LayerNorm	LayerNorm	2048	82.0	45.3	ViT default
RMSNorm	RMSNorm	2048	82.2	45.1	LN to RMS
MH-RMSNorm	RMSNorm	2048	84.0	37.3	partial
RMSNorm	MH-RMSNorm	2048	84.0	37.2	partial
MH-RMSNorm	MH-RMSNorm	2048	**84.2**	**36.9**	ours

Table 6: Ablation on where to apply MH-RMSNorm. We replace normalization layers in either the attention branch, the MLP branch, or both, while keeping all other settings unchanged. Results are reported at width 2048 with depth fixed to 12, using ImageNet-1K top-1 accuracy and ImageNet-256 FID. The full replacement (ours) performs best, while partial replacement already provides most of the gain.

Setting	Top-1 (%) ↑	FID ↓
RMSNorm for all layers	82.2	45.1
MH-RMSNorm for only shallow layers (layers 1-4)	83.1	41.0
MH-RMSNorm for only middle layers (layers 5-8)	83.7	39.6
MH-RMSNorm for only deep layers (layers 9-12)	83.9	38.4
MH-RMSNorm for all layers	**84.2**	**36.9**

Table 7: Ablation on which layers use MH-RMSNorm. We apply MH-RMSNorm to different depth ranges of a 12-block transformer while keeping all other settings unchanged. Top-1 is evaluated on ImageNet-1K and FID is evaluated on ImageNet-256. Our default setup is highlighted in orange.

of outlier activations can dominate the normalization statistics, suppressing the remaining dimensions and attenuating gradients, especially under low-precision training. To address this, we propose Multi-Head RMSNorm (MH-RMSNorm), a simple drop-in replacement that partitions the feature dimension into moderate-sized chunks and normalizes each chunk independently, preventing global domination while preserving computational efficiency. Across standard benchmarks, MH-RMSNorm consistently improves ImageNet-1K classification and ADE20K segmentation, with gains that become more apparent as model width increases. Beyond these settings, a dedicated width-scaling study shows that conventional global normalizers become unreliable in the ultra-wide regime, whereas MH-RMSNorm enables stable scaling up to very large widths and maintains strong performance on both discriminative and diffusion-based generative tasks. Ablations further reveal that the method is robust to the choice of head count within a broad range, that the benefit generalizes across normalization families but is strongest with RMS-based statistics, and that applying MH-RMSNorm throughout the network provides the best results, with deeper layers being particularly important. Overall, MH-RMSNorm offers a practical normalization primitive for training and scaling wide transformers, turning width from a source of instability into a reliable axis for increasing model capacity.

Acknowledgements

This work was partially funded by an internal grant from the Johns Hopkins University. The authors FW and AY are funded by ONR N000142512132.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016)
2. Bachlechner, T., Majumder, B.P., Mao, H.H., Cottrell, G., McAuley, J.: Rezero is all you need: Fast convergence at large depth. In: ICLR (2021)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report (2023)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
5. Chen, T., Zhang, Z., Cheng, Y., Awadallah, A., Wang, Z.: The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. CVPR (2022)
6. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. ICLR (2024)
7. Ding, M., Yang, Z., Hong, W., et al.: Cogview: Mastering text-to-image generation via transformers. In: NeurIPS (2021)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
9. Graef, N., Clapp, M., Wasielewski, A.: Flash normalization: fast rmsnorm for LLMs (2024)
10. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models (2024)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
12. Henry, A., Dachapally, P.R., Pawar, S., Chen, Y., Mueller, J., Koyejo, S.: Query-key normalization for transformers. In: EMNLP (2020)
13. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: ECCV (2016)
14. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer v2: Scaling up capacity and resolution. In: CVPR (2022)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
16. Mohan, N.V., Zhao, Y., Welleck, S., Zhang, Y., Mao, Y., Tian, Y., Wang, Z.: Gated removal of normalization in transformers enables stable training and efficient inference (2026)
17. Nguyen, T.Q., Salazar, J.: Transformers without tears: Improving the normalization of self-attention. IWSLT (2019)
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve,

G., Xu, H., Jégou, H., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. TMLR (2024)
19. Oquab, M., Darcet, T., Moutakanni, T., et al.: Dinov3 (2025)
20. Qi, X., Wang, J., Chen, Y., Shi, Y., Zhang, L.: Lipsformer: Introducing lipschitz continuity to vision transformers. In: ICLR (2023)
21. Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., et al.: Qwen2.5 technical report (2024)
22. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020)
23. Shen, S., Yao, Z., Gholami, A., Mahoney, M., Keutzer, K.: Powernorm: Rethinking batch normalization in transformers. In: ICML (2020)
24. Shleifer, S., Weston, J., Ott, M.: Normformer: Improved transformer pretraining with extra normalization. ICLR (2022)
25. Sun, M., Chen, X., Kolter, J.Z., Liu, Z.: Massive activations in large language models. COLM (2024)
26. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
27. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Deit iii: Revenge of the vit. ECCV (2022)
28. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: ICCV (2021)
29. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., et al.: LLaMA: Open and efficient foundation language models (2023)
30. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models (2023)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
32. Wang, F., Ren, S., Zhang, T., Neskovic, P., Bhattad, A., Xie, C., Yuille, A.: ViT-5: Vision transformers for the mid-2020s (2026)
33. Wang, H., Ma, S., Dong, L., Wei, F., Liu, T.Y.: Deepnet: Scaling transformers to 1,000 layers. PAMI (2024)
34. Wu, Y., He, K.: Group normalization. In: ECCV (2018)
35. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV (2018)
36. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.Y.: On layer normalization in the transformer architecture. ICML (2020)
37. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., et al.: Qwen3 technical report (2025)
38. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., et al.: Qwen2 technical report (2024)
39. You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Hsieh, C.J.: Large batch optimization for deep learning: Training BERT in 76 minutes. In: ICLR (2020)
40. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICML (2023)
41. Zhang, B., Sennrich, R.: Root mean square layer normalization (2019)

42. Zhang, H., Dauphin, Y.N., Ma, T.: Fixup initialization: Residual learning without normalization. In: ICLR (2019)
43. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV (2019)
44. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer (2021)