CLUSTERING IN SPIKING TRANSFORMERS

Tristan Peat
Georgia Institute of Technology
Atlanta, GA 30313, USA
tpeat@gatech.edu

ABSTRACT

Transformers have revolutionized deep learning, yet the mechanisms behind their exceptional performance remain poorly understood. We investigate the hypothesis that Transformers function as clustering machines by comparing traditional Vision Transformers (ViT) and Spiking Transformers (Spikformer) through the lens of Wasserstein gradient flows in particle systems. Our analysis reveals that while traditional Transformers maintain independent attention heads, Spiking Transformers show progressive clustering in deeper layers. Through ablation studies, we demonstrate how skip connections and feed-forward networks critically influence this clustering behavior. We develop theoretical foundations for understanding membrane potential dynamics in Spiking Transformers and provide insights into the design principles of both architectures. These results advance our understanding of how different transformer variants process and organize information through their network depths.

1 INTRODUCTION

Transformers have revolutionized deep learning through their ability to capture long-range dependencies in sequential data (Vaswani et al., 2023). Recent theoretical work has revealed that the self-attention mechanism, when viewed through the lens of optimal transport, exhibits intriguing clustering behavior. Specifically, Geshkovski et al. (2024a) demonstrated that tokens in pretrained Transformer architectures tend to increase in similarity with network depth, ultimately converging to a single cluster as training time approaches infinity.

This clustering phenomenon bears resemblance to the neural collapse observed by Papyan et al. (2020), though the underlying mechanisms differ. The self-attention mechanism can be viewed as an interacting particle system, where tokens act as particles governed by a vector field derived from their pairwise interactions. This perspective has proven valuable in understanding the convergence properties of Transformer architectures.

Parallel to these theoretical developments, recent work has explored the integration of biological inspiration into Transformer architectures through Spiking Neural Networks (SNNs). The Spikformer architecture (Zhou et al., 2022) represents a notable advancement in this direction, replacing traditional floating-point operations with discrete spike sequences and incorporating leaky integrate-and-fire (LIF) neuron dynamics.

Our work investigates the intersection of these two research directions. We aim to understand how the introduction of spiking dynamics affects the clustering behavior observed in traditional Transformers. Specifically, we examine the training dynamics of vanilla Transformers versus Spiking Transformers, with particular attention to the role of post-attention MLPs, skip-connections, and attention head correlations across network depth. Furthermore, we develop theoretical foundations for understanding how dynamic membrane potentials in Spiking Transformers influence their clustering properties.

2 RELATED WORKS

2.1 THE TRANSFORMER AND SELF-ATTENTION

Transformers (Vaswani et al., 2023) have revolutionized deep learning through their self-attention mechanism, which captures relationships between inputs by computing compatibility scores between queries and keys. Each input token x_i generates a query $Q(x_i)$ that interacts with keys $K(x_j)$ via inner products, with high compatibility scores determining how much of value $V(x_j)$ contributes to the final representation. Recent theoretical work has shown that isolated self-attention converges to rank-1 solutions (Dong et al., 2023), though skip connections prevent this collapse. Token similarity tends to increase with network depth (Geshkovski et al., 2024a), but practical constraints often lead to convergence at saddle points rather than single clusters (Sander et al., 2022).

2.2 SPIKING NEURAL NETWORKS AND THE LIF MODEL

Spiking Neural Networks (SNNs) represent the third generation of neural architectures (Maass, 1997), using discrete spike sequences instead of continuous values for improved energy efficiency. The fundamental unit is the Leaky Integrate-and-Fire (LIF) neuron, which accumulates membrane potential according to:

$$V(t+1) = V(t) + \frac{1}{C}(I(t) - g_L(V(t) - V_{\text{rest}})) \tag{1}$$

When the potential exceeds a threshold, the neuron fires and resets. This binary threshold operation requires surrogate gradients for training (Wu et al., 2018).

2.3 SPIKING TRANSFORMERS

The Spikformer architecture (Zhou et al., 2022) merges transformers with SNNs through Spiking Self-Attention (SSA), replacing traditional floating-point operations with spike-based computations. This approach leverages natural sparsity and binary activations, introducing temporal dynamics through membrane potential accumulation. Recent variants like Spike-Driven Transformers (Yao et al., 2023) further optimize the architecture for efficiency, though their impact on clustering dynamics remains unexplored.

3 DEFINING THE DYNAMICS

3.1 SELF-ATTENTION AS A PARTICLE SYSTEM

We can formulate Transformer self-attention as an interacting particle system where input tokens $\{x_1 \cdots x_n\} \in \mathbb{R}^d$ represent particles. Due to positional encoding, we view these as a probability measure:

$$\{x_1 \cdots x_n\} \in \mathbb{R}^d \iff \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \tag{2}$$

The standard attention mechanism (Vaswani et al., 2023) is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

This naturally leads to mean-field dynamics where each particle evolves according to:

$$x_i(t) = X_t[\mu(t)](x_i(t)) \tag{4}$$

Following Geshkovski et al. (2024b), we can simplify this by fixing $V = I_d$, $Q^T K = \beta I_d$ and projecting onto the unit sphere to obtain:

$$\dot{x}(t) = \text{Proj}_{\tau_{x_i(t)}S^{d-1}} \frac{\sum_{j=1}^{n} x_j(t) e^{\beta\langle x_i(t), x_j(t)\rangle}}{\sum_{k=1}^{n} e^{\beta\langle x_i(t), x_k(t)\rangle}} \tag{5}$$

Full derivations and extensions to include MLPs are provided in Appendix Sections A.1 and A.2.

3.2 SPIKING SELF ATTENTION AS A PARTICLE SYSTEM

We consider the Spiking Self-Attention (SSA) mechanism introduced by (Zhou et al., 2022). Let $\mathcal{X} = x_i(t)_{i=1}^N \subset \mathbb{R}^d$ represent N particles evolving in discrete time steps. Each particle undergoes three main transformations: First, the projection onto the unit sphere:

$$\tilde{q}_i(t) = \text{Proj}\mathbb{S}^{d-1}(W_q x_i(t)), \quad \tilde{k}_i(t) = \text{Proj}\mathbb{S}^{d-1}(W_k x_i(t)), \quad \tilde{v}_i(t) = \text{Proj}\mathbb{S}^{d-1}(W_v x_i(t)) \tag{6}$$

Second, these projections pass through Leaky Integrate-and-Fire (LIF) dynamics:

$$\text{LIF}(u(t)) = \begin{cases} 1, & \text{if } U(t) \geq v_{\text{threshold}} \; 0, \\ \text{otherwise} \end{cases} \tag{7}$$

Finally, the particle evolution follows:

$$x_i(t+1) = \text{Proj}\mathbb{S}^{d-1}\left(\text{LIF}\left(W_p\left(\sum_{j=1}^{N} \alpha_{ij}(t) v_j(t)\right)\right)\right) \tag{8}$$

where $\alpha_{ij}(t) = \text{LIF}(\frac{1}{8}\langle q_i(t), k_j(t)\rangle)$ represents the attention weights. To analyze the continuous limit, we replace the discrete LIF with a continuous approximation $\sigma_\tau(u) = \frac{1}{1+e^{-\tau(u-v_{\text{threshold}})}}$. Next we follow Geshkovski et al. (2024b) by simplifying weight matrices $W_q = W_k = W_v = W_p = \beta I_d$. This yields the mean-field dynamics:

$$\dot{x}(t) = \text{Proj}_{\tau_{x(t)}}\mathbb{S}^{d-1}\left(\sigma_\tau\left(\frac{\int_{\mathbb{S}^{d-1}} y e^{\beta\langle x(t), y\rangle}\mu(t)(dy)}{\int_{\mathbb{S}^{d-1}} e^{\beta\langle x(t), y\rangle}\mu(t)(dy)}\right)\right) \tag{9}$$

where β controls interaction strength and $\mu(t)$ represents the limiting empirical measure.

4 PROBLEM STATEMENT

We hypothesize that Transformers are clustering machines and wish to expand on the work of Geshkovski et al. (2024a) in two ways:

1. Analyze the training dynamic differences between a vanilla ViT and a spiking Transformer, investigating the importance of post attention MLP, necessity of skip-connections, and correlation between attention heads with depth
2. Design a theoretical understanding of how dynamic (learned) membrane potentials impact the clustering in spiking Transformers

5 METHODOLOGY

We performed a systematic ablation study by creating three variants of both ViT and Spikformer architectures: (1) baseline models retaining all components, (2) models without MLP blocks, and (3) models with skip connections removed. This design enables direct comparison of component importance across traditional transformers and spike-based architectures. The baseline models maintain full architectural complexity, while MLP removal substantially reduces parameters and computational requirements as shown in Table 1. Skip connection removal has minimal impact on model size but allows us to study their role in gradient flow and training dynamics.

5.1 ARCHITECTURE OVERVIEW

The ViT serves as our traditional Transformer baseline, following the original architecture where input images are divided into fixed-size patches, linearly embedded, and processed through a sequence of transformer blocks. Each block contains a multi-head self-attention mechanism followed by a multi-layer perceptron (MLP), with layer normalization and skip connections throughout. This encoder-only architecture culminates in a classification head for direct prediction.

In contrast, Spikformer adapts the transformer architecture to operate with spiking neurons, introducing temporal dynamics through discrete time steps. It replaces the standard patch embedding with a Spiking Patch Splitting (SPS) module that uses hierarchical convolutional layers with integrated spiking neurons. The self-attention mechanism is modified to work with spike-based representations, incorporating LIF neurons with carefully tuned membrane potential decay. This architecture maintains the general structure of ViT while fundamentally changing how information is processed and propagated through the network.

5.2 Exploring Training Dynamics Methodology

We trained all models on the CIFAR-10 dataset using identical hyperparameters to ensure fair comparison, as detailed in Table 4. The architecture follows a standard ViT design with an embedding dimension of 384, a 4x4 patch size, and 12 layers, each with 6 attention heads.

For optimization, we employed AdamW with an initial learning rate of 5e-4 and weight decay of 6e-2. The learning rate followed a cosine decay schedule with a 3-epoch linear warmup period, which helped stabilize early training. The models were trained for 83 epochs with a batch size of 128, leveraging mixed precision acceleration on a single NVIDIA V100-32GB GPU for roughly 4 hours per model variant.

5.3 Computational Efficiency Analysis

Removing skip connections has negligible impact on model complexity (0.06% parameter reduction). However, eliminating MLP blocks significantly reduces both parameters (66.5%) and FLOPs in both architectures (Table 1). The ViT model shrinks from 21.34M to 7.14M parameters, while Spikformer reduces from 23.58M to 9.35M, suggesting that MLP optimization could yield substantial efficiency gains while preserving performance.

Model Variant	**Parameters**	**FLOPs**
Baseline ViT	21.34M	1.38B
ViT No MLP	**7.14M**	**462.54M**
ViT No Skip Connections	21.33M	1.38B
Baseline Spikformer	23.58M	7.36B
Spikformer No MLP	**9.35M**	**3.71B**
Spikformer No Skip Connections	23.58M	7.36B

Table 1: Computational complexity comparison across model variants

5.4 Attention Head Correlation

Following (Agarwal et al., 2024), we analyze head correlations across network depths to understand information distribution patterns. For ViT models, we measure correlations between scaled dot-product attention outputs directly. For Spikformers, we first average attention patterns across time steps to enable fair comparison while preserving spike-based characteristics. Correlation coefficients between attention heads are computed for each network layer using standard statistical measures of covariance and standard deviation.

6 Results

6.1 Performance Analysis

The complete training trajectories for accuracy and loss metrics are provided in Section C of the Appendix (Figures 3–6). These graphs reveal several important patterns in the training dynamics. The baseline ViT shows consistent and stable improvement in both accuracy and loss throughout training (Figures 3a, 4a), while the Spikformer exhibits more volatile learning behavior, particularly in early epochs (Figures 3b, 4b). Models without MLPs (Figure 5) demonstrate slower convergence compared to their baseline counterparts, though they eventually achieve reasonable performance.

Most notably, the variants without skip connections (Figure 6) show minimal learning progress, with both architectures struggling to move beyond random-level performance, as evidenced by their consistently high loss values and poor accuracy metrics.

6.2 Final Epoch Performance

Table 2 presents the performance metrics at the final training epoch. The baseline Vision Transformer (ViT) demonstrated superior performance with 84.94% top-1 accuracy, significantly outperforming other variants. Notably, the removal of MLPs resulted in a moderate performance degradation (75.49% top-1 accuracy), while the removal of skip connections led to a dramatic performance collapse (10.00% top-1 accuracy).

Model	Train Loss	Val Loss	Top-1 (%)	Top-5 (%)
Baseline ViT	1.2136	0.5602	84.94	99.04
ViT No MLP	1.4025	0.8055	75.49	97.61
ViT No Skip Connections	2.3026	2.3027	10.00	50.00
Baseline Spikformer	2.0030	1.6754	41.46	88.75
Spikformer No MLP	2.0872	1.8528	34.65	84.82
Spikformer No Skip Connections	2.2897	2.2519	15.68	66.14

Table 2: Final epoch performance comparison across different models

6.3 Best Performance Analysis

Table 3 shows the best performance achieved by each model during training. The baseline ViT maintained its superior performance, while the Spikformer architecture showed notable improvement in its best-case scenario compared to its final epoch performance.

Model	Train Loss	Val Loss	Top-1 (%)	Top-5 (%)
Baseline ViT	1.2136	0.5602	84.94	99.04
ViT No MLP	1.3876	0.8055	75.53	98.00
ViT No Skip Connections	2.3026	2.3027	10.00	50.00
Baseline Spikformer	1.6515	1.0474	67.29	97.45
Spikformer No MLP	1.6056	1.0067	**68.42**	**97.50**
Spikformer No Skip Connections	2.2764	2.2265	17.49	70.88

Table 3: Best performance comparison across different models

6.4 Spiking Self-Attention Clustering Analysis

To investigate the clustering behavior of Spiking Self-Attention (SSA), we performed a detailed analysis of particle dynamics under various parameter configurations. We focused on two key parameters: the temperature β controlling the strength of particle interactions, and the activation steepness τ determining the sharpness of the spiking behavior.

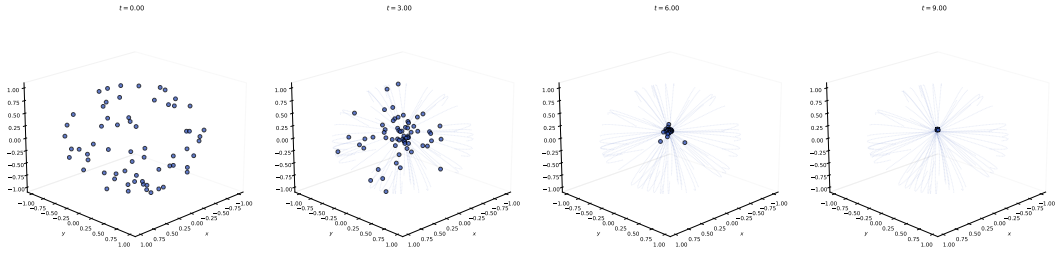
6.4.1 Experimental Setup

We initialized $N = 64$ particles on the unit sphere in $\mathbb{R}^3$ and evolved them according to the SSA dynamics described in Section 3.3. To capture the full range of behaviors, we tested multiple parameter combinations for temperature $\beta \in \{0.5, 1.0, 2.0\}$ and activation steepness $\tau \in \{1.0, 2.0, 4.0\}$.

6.4.2 Clustering Dynamics

Figure 1 shows the evolution of particle positions at four different time points for $\beta = 1.0$ and $\tau = 2.0$. The visualization reveals an initial dispersion phase where particles explore the sphere

surface, then a formation of intermediate clusters as particles begin to influence each other then progressive merging of clusters over time, and finally convergence to stable cluster configurations.



(a) $t = 0.0$ (b) $t = 3.0$ (c) $t = 6.0$ (d) $t = 9.0$

Figure 1: Evolution of particle clustering on the unit sphere under Spiking Self-Attention dynamics with $\beta = 1.0$ and $\tau = 2.0$. Particles (blue dots) are shown at four time points, with dashed lines indicating their trajectories. The progression shows initial dispersion followed by gradual cluster formation.

6.5 Attention Head Correlation Analysis

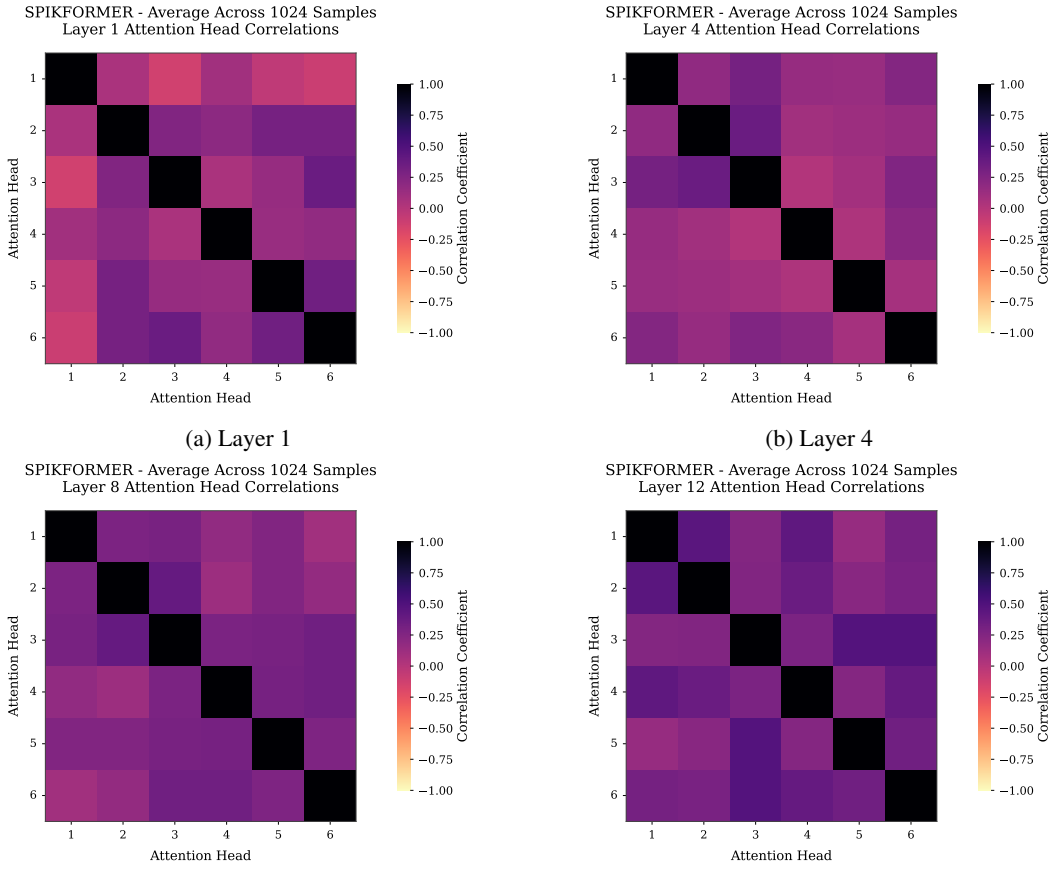
Our analysis reveals fundamentally different attention patterns between ViT and Spikformer architectures. Spikformer exhibits increasing head correlation with network depth (Figure 2), indicating progressive information clustering. In contrast, ViT maintains or decreases head correlation across layers (Figure 7), preserving independent feature extraction capabilities. This distinction helps explain the performance gap between architectures: ViT's independent heads enable diverse feature extraction, while Spikformer's correlation tendency may limit its representational capacity. These findings provide empirical evidence for theoretical predictions about attention dynamics while highlighting key differences between traditional and spike-based mechanisms.

7 Discussion and Limitations

Our analysis reveals fundamentally different clustering behaviors between ViT and Spikformer architectures. While ViTs maintain independent attention heads and achieve superior accuracy (84.94% vs 41.46% top-1 on CIFAR-10), Spikformers exhibit progressive clustering with depth, potentially limiting their representational capacity. The dramatic performance collapse when removing skip connections (to 10.00% and 15.68% accuracy respectively) demonstrates their crucial role in preventing excessive clustering and maintaining distinct information pathways. Several limitations constrain our findings. Our experiments used only CIFAR-10 with modest model sizes, and computational constraints prevented extensive hyperparameter optimization. The temporal dynamics of Spikformer clustering behavior and convergence properties remain partially unexplored. Additionally, our implementation-specific conclusions may not generalize across all architectural variants.

8 Conclusion

This work advances our understanding of transformer clustering dynamics, suggesting that excessive clustering may impair rather than enhance performance. While spike-based processing offers efficiency benefits, maintaining independent feature extraction capabilities appears crucial for optimal performance. These insights inform the design of future neuromorphic architectures, highlighting the delicate balance between information mixing and preservation in deep networks. Future research should explore larger-scale datasets, detailed temporal dynamics, and broader architectural variants to further validate and expand these findings.



(a) Layer 1

(b) Layer 4

(c) Layer 8

(d) Layer 12

Figure 2: Attention head correlation patterns across different layers in Spikformer. The heatmaps show correlation coefficients between attention heads, with darker colors indicating stronger correlations. Layer 1 (a) shows initial head independence, while deeper layers (b-d) demonstrate increasing correlation patterns, suggesting progressive information clustering through the network depth.

REFERENCES

Saurabh Agarwal, Bilge Acun, Basil Hosmer, Mostafa Elhoushi, Yejin Lee, Shivaram Venkataraman, Dimitris Papailiopoulos, and Carole-Jean Wu. Chai: Clustered head attention for efficient llm inference, 2024. URL https://arxiv.org/abs/2403.08058.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023. URL https://arxiv.org/abs/2103.03404.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics, 2024a. URL https://arxiv.org/abs/2305.05465.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers, 2024b. URL https://arxiv.org/abs/2312.10794.

Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(97)00011-7. URL https://www.sciencedirect.com/science/article/pii/S0893608097000117.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):

24652–24663, September 2020. ISSN 1091-6490. doi: 10.1073/pnas.2015509117. URL http://dx.doi.org/10.1073/pnas.2015509117.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention, 2022. URL https://arxiv.org/abs/2110.11773.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience*, 12, May 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00331. URL http://dx.doi.org/10.3389/fnins.2018.00331.

Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer, 2023. URL https://arxiv.org/abs/2307.01694.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer, 2022. URL https://arxiv.org/abs/2209.15425.

A Mathematical Grounds

A.1 Formal Definition of Transformer Block

A transformer block is a parameterized function class $f_\theta : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$. If $x \in \mathbb{R}^{n\times d}$, then $f_\theta(x) = z$, where

$$Q^{(h)}(x_i) = W_{h,q}^T x_i, \quad K^{(h)}(x_i) = W_{h,k}^T x_i, \quad V^{(h)}(x_i) = W_{h,v}^T x_i, \quad W_{h,q}, W_{h,k}, W_{h,v} \in \mathbb{R}^{d\times k} \tag{1}$$

$$\alpha_{i,j}^{(h)} = \text{softmax}_j\left(\frac{\langle Q^{(h)}(x_i), K^{(h)}(x_j)\rangle}{\sqrt{k}}\right) \tag{2}$$

$$u_i' = \sum_{h=1}^{H} W_{c,h}^T \sum_{j=1}^{n} \alpha_{i,j}^{(h)} V^{(h)}(x_j), \quad W_{c,h} \in \mathbb{R}^{k\times d} \tag{3}$$

$$u_i = \text{LayerNorm}(x_i + u_i'; \gamma_1, \beta_1), \quad \gamma_1, \beta_1 \in \mathbb{R}^d \tag{4}$$

$$z_i' = W_2^T \text{ReLU}(W_1^T u_i), \quad W_1 \in \mathbb{R}^{d\times m}, \quad W_2 \in \mathbb{R}^{m\times d} \tag{5}$$

$$z_i = \text{LayerNorm}(u_i + z_i'; \gamma_2, \beta_2), \quad \gamma_2, \beta_2 \in \mathbb{R}^d \tag{6}$$

A.2 Potential dynamics of 2-layer MLP

Preliminary equation for the dynamics with an MLP can be given by:

$$\dot{z_i}(t) = W\sigma\left(\sum_{j=1}^{n}\left(\frac{e^{\langle Qe^{tV}z_i(t), Ke^{tV}z_j(t)\rangle}}{\sum_{k=1}^{n} e^{\langle Qe^{tV}z_j(t), Ke^{tV}z_k(t)\rangle}}\right)(z_j(t) - z_i(t))\right)$$

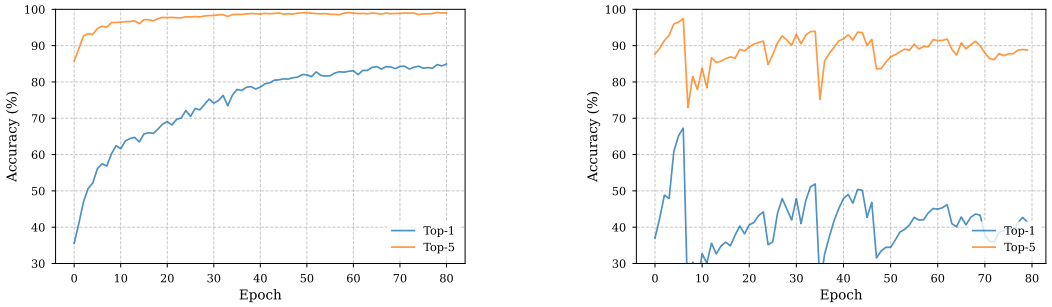
for $i \in [n]$ and $t \geq 0$. We can also include a bias vector $b \in \mathbb{R}^d$ either inside or outside of the activation function to allow for translations

B TRAINING HYPERPARAMETERS

Hyperparameter	**Value**
Architecture	
Image Size	32×32
Patch Size	4×4
Embedding Dimension	384
Transformer Depth	12
Number of Heads	6
MLP Ratio	4
Number of Classes	10
Training	
Optimizer	AdamW
Learning Rate	5e-4
Weight Decay	6e-2
Batch Size	128
Epochs	83
Warmup Epochs	3
Scheduler	Cosine
Label Smoothing	0.1
Augmentation	
Mixup α	0.5
Mixup Probability	1.0
CutMix	0.0
Normalization	
Mean	[0.4914, 0.4822, 0.4465]
Std	[0.2470, 0.2435, 0.2616]

Table 4: Complete training configuration and model hyperparameters.

C TRAINING ACCURACY AND LOSS GRAPHS

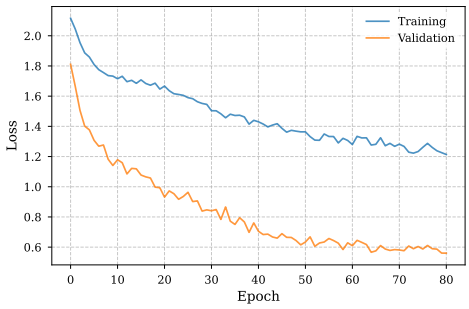


(a) Baseline ViT Accuracy

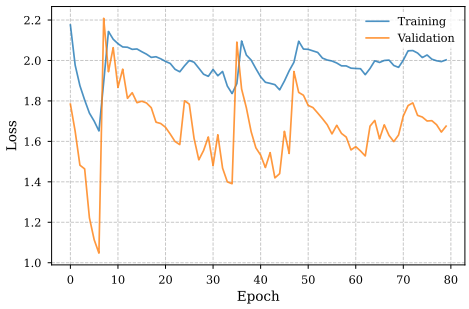
(b) Baseline Spikformer Accuracy

Figure 3: Baseline Accuracies

D ATTENTION HEAD CORRELATION IN VIT

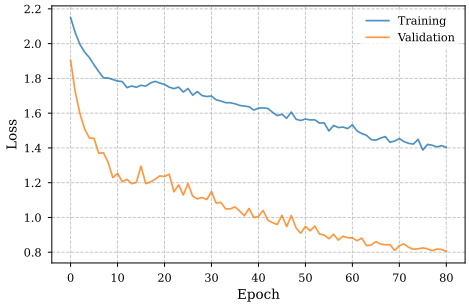


(a) Baseline ViT Loss

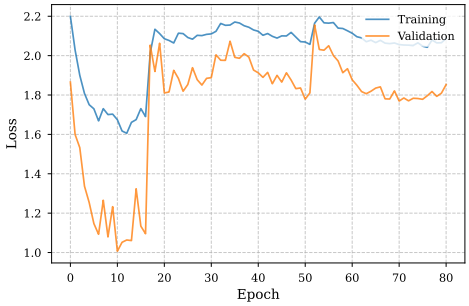


(b) Baseline Spikformer Loss

Figure 4: Baseline losses

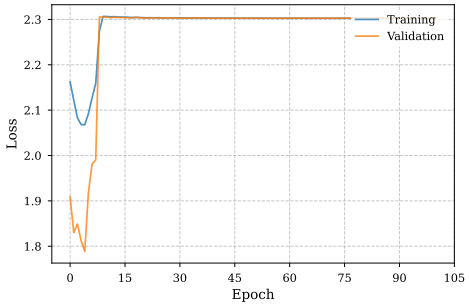


(a) ViT No MLP Loss

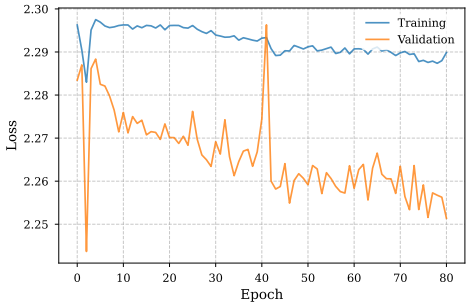


(b) Spikformer No MLP Loss

Figure 5: No MLP losses

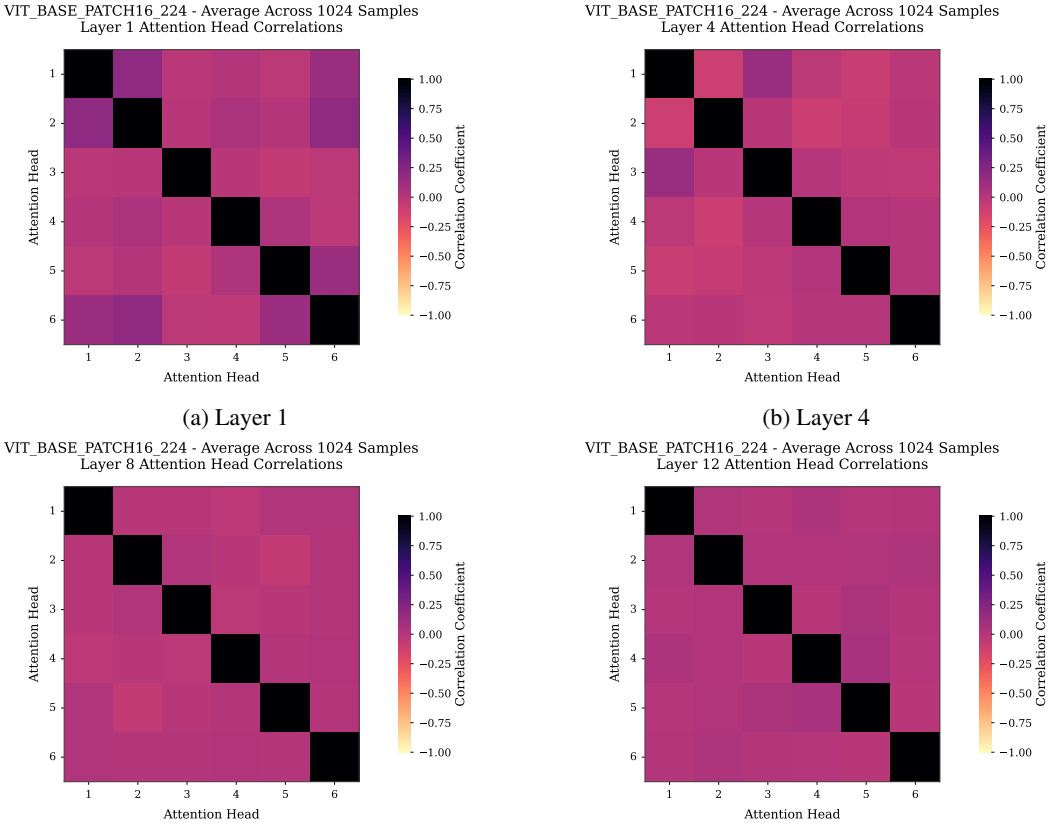


(a) ViT No Skip Connections Loss



(b) Spikformer No Skip Connections Loss

Figure 6: No Skip losses



(a) Layer 1

(b) Layer 4

(c) Layer 8

(d) Layer 12

Figure 7: Attention head correlation patterns across different layers in ViT. Surprisingly, the heat map doesn't show a strong evolution of the correlation. Actually showing a decrease in correlation between attention heads with depth.