IMPROVED BELIEF-ATTENTION IN VISION TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, Belief-Attention (Zhang, 2025) was proposed to improve Transformer performance by orthogonally projecting the softmax-weighted sum of V vectors relative to the original V vectors and retaining the perpendicular component as a residual signal. In this paper, we first present an ablation study demonstrating that the projected component also encodes valuable token correlation information that should not be overlooked. Based on this insight, we propose Belief2-Attention, an extension that effectively leverages both components. Specifically, the projected component passes through an activation function and a linear mapping before being merged with the target token—effectively forming an internal two-layer feedforward network (FFN) within the attention mechanism. Furthermore, while standard attention models token correlations solely via the inner product $QK^\top$, we modify this formulation to $(Q+K)K^\top$ to enrich pairwise token representations for vision tasks. This allows $KK^\top$ to capture symmetric correlations, leaving $QK^\top$ to model complementary directional dependencies. We demonstrate that Belief2-Attention possesses greater expressive capacity than standard attention under mild conditions. Finally, we validate the effectiveness of Belief2-Attention on vision tasks, including image classification and image segmentation.

1 INTRODUCTION

In the last decade, Transformers (Vaswani et al., 2017) have driven major progress across diverse areas of data analysis, including natural language processing (NLP) (Achiam et al., 2023; Touvron et al., 2023), computer vision (Dosovitskiy et al., 2021), image generation and editing (Peebles & Xie, 2023; Hatamizadeh et al., 2024; Zhang et al., 2023), and audio processing (Latif et al., 2023). At the core of these models lies the attention mechanism, which allows them to capture long-range dependencies within sequences. This is achieved by forming a weighted sum of value (V) vectors, where the weights are derived by computing the inner-product between the query (Q) and key (K) vectors, followed by the softmax operation. In essence, attention enables each token to selectively incorporate information from all other tokens. After the attention step, a feedforward network (FFN) operates on each token independently, serving as a form of localized information processing. More recently, large language models (LLMs) exploit a so-called mixture of experts (MoE) as an extension of basic FFN to improve the performance, where at the inference stage, only certain percentage of weights in the FFN layer are activated depending on the particular input.

One prominent research direction focuses on reducing the quadratic computational complexity inherent in the standard attention layer when processing long token sequences. Various simplified attention schemes have been proposed, which include, for example, LinFormer (Wang et al., 2020), LongFormer (Beltagy et al., 2020), ReFormer (Kitaev et al., 2020), FlashAttention (Dao, 2023), RingAttention (Liu et al., 2023), BurstAttention (Sun et al., 2023). FlashAttention is being widely used in practical situations as it reduces the computational complexity considerably without introducing any approximation in the standard attention layer.

Another important research trend attempts to modify the attention layer in Transformer for better performance rather than for reduction of computational complexity. For instance, the work (Zhang, 2025) proposes a so-called Belief-Attention (see Section 2 for details). The basic idea is to first perform an orthogonal projection of the softmax-based weighted sum of V vectors onto the original V vectors and then taking the perpendicular component as the residual signal. On the contrary, the standard attention takes the weighted sum of V vectors directly as the residual signal. It is argued in (Zhang, 2025) that the perpendicular component reflects the discrepancy between the

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.

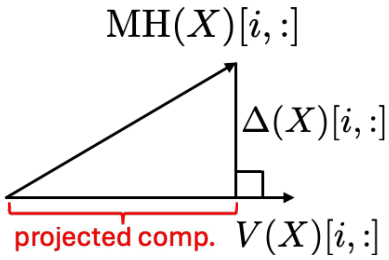


Figure 1: Orthogonal projection.

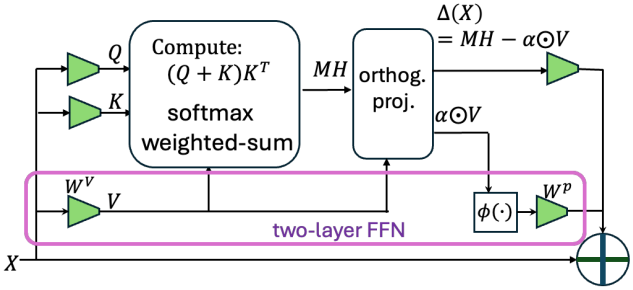


Figure 2: Demonstration of Belief2-Attention, where MH refers to "multi-head attention".

weighted sum of V vectors and the V vectors themselves, causing the tokens to update primarily along the tangent directions rather than according to their magnitudes. The recent work (Zhai, 2026) also recommends to take the perpendicular component as the residual signal in an attention layer. The main difference between (Zhang, 2025) and (Zhai, 2026) is that the former considers both per-attention-head and global perpendicular components, while the latter only considers the per-attention-head perpendicular component (see Remark 1 later on for a detailed discussion).

In this work, we first perform an ablation study on Belief-Attention, demonstrating that the projected component also encodes valuable token correlation information that should not be discarded. Based on this observation, we propose passing the projected component through an activation function (which may be an identity mapping) and a linear layer before merging it with the target token sequence. From a high-level viewpoint, this attention block incorporates an implicit two-layer FFN to process the projected component (Fig. 2). Additionally, because spatial interactions in vision tasks are non-causal, we modify the inner-product term from $QK^\top$ to $(Q+K)K^\top$, enabling $KK^\top$ and $QK^\top$ to capture symmetric and directional correlations, respectively. We refer to this mechanism as *Belief2-Attention*.

Empirical evaluations on image classification and segmentation demonstrate that Belief2-Attention yields substantial performance gains over standard attention. In contrast, Belief-Attention struggles on tasks such as image segmentation due to the omission of the projected component. To maintain parameter parity across baselines, the dimensions of conventional FFN layers were scaled down accordingly. Furthermore, extensive ablation studies reveal that even after removing all standard FFN layers, a model relying solely on stacked Belief2-Attention layers achieves performance comparable to standard Vision Transformers.

2 Ablation Study of Belief-Attention

In this section, we first briefly revisit Belief-Attention in (Zhang, 2025). We then perform an ablation study of Belief-Attention for image classification over CIFAR100. We demonstrate experimentally that the projected component after orthogonal projection still carries noticeable information about token correlations, which should not be ignored by Belief-Attention.

2.1 Revisiting Belief-Attention

As mentioned earlier, Belief-Attention includes both the per-attention-head and global perpendicular components as the residual signals before being processed by linear mappings. Following the convention of python-based implementation of attention, suppose a tensor $X \in \mathbb{R}^{n\times d}$ of n tokens is the input from the layer below in a Transformer, where each token is of dimension d. We use the row vector $X[i,:]$ to denote the ith token. The update expression of Belief-Attention can be represented as (see (Zhang, 2025; Dosovitskiy et al., 2021))

$$\mathrm{H}_m(X) = \mathrm{attention}(\overbrace{XW_m^Q}^{Q_m}, \overbrace{XW_m^K}^{K_m}, \overbrace{XW_m^V}^{V_m}) \tag{1}$$

$$\Delta_m(X)[i,:] = \underbrace{H_m(X)[i,:] - \overbrace{\beta_{m,i}V_m[i,:]}^{\text{projected comp.}}}_{\text{perpendicular comp.}} \quad \beta_{m,i} = \frac{\langle \mathrm{H}_m(X)[i,:], V_m[i,:]\rangle}{\langle V_m[i,:], V_m[i,:]\rangle} \tag{2}$$

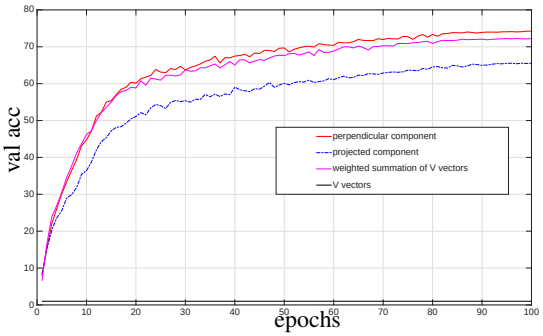


Figure 3: Impact of different residual signals in the attention layers of a Transformer being trained over CIFAR100 for image classification. The first open-source in Table 7 is utilized for this experiment.

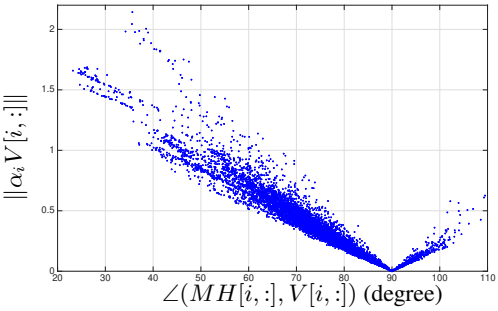


Figure 4: Scatter plot of the angle $\angle(MH[i,:], V[i,:])$ versus $\|\alpha_i V[i,:]\|$ for each token index i in the 8th transformer block after training the standard ViT model over CIFAR100. (check Appendix A).

$$\text{MH}(X) = \text{Concat}(\text{H}_1(X), \ldots, \text{H}_M(X)) \tag{3}$$

$$V(X) = \text{Concat}(V_1, \ldots, V_M) \tag{4}$$

$$\Delta^l(X) = \text{Concat}(\Delta_1(X), \ldots, \Delta_M(X)) \tag{5}$$

$$\Delta(X)[i,:] = \underbrace{\text{MH}(X)[i,:] - \overbrace{\alpha_i V(X)[i,:]}^{\text{projected comp.}}}_{\text{perpendicular comp.}} \quad \alpha_i = \frac{\langle \text{MH}(X)[i,:], V(X)[i,:] \rangle}{\langle V(X)[i,:], V(X)[i,:] \rangle} \tag{6}$$

$$X \Leftarrow X + \Delta(X)W + \Delta^l(X)W^l, \tag{7}$$

where pre-normalization is omitted for simplicity, $i \in \{1, 2, \ldots, n\}$ denotes the ith token, $m \in \{1, 2 \ldots, M\}$ denotes the mth attention-head, (W_m^Q, W_m^K, W_m^V) are the three learnable matrices for computing $(Q_m, K_m, V_m) \in (\mathbb{R}^{n \times d_m}, \mathbb{R}^{n \times d_m}, \mathbb{R}^{n \times d_m})$ for the mth attention. The notations H and MH stand for "head" and "multi-head", respectively.

By inspection of (1)-(7), we can conclude that $\Delta_m(X)$ is the local perpendicular component for the mth attention head and Δ is the global one after projecting each $MH(X)[i,:]$ onto its associated $V(X)[i,:]$ vector in a token-wise manner. To facilitate discussion later on, we rewrite (2) and (6) in a compact form as:

$$\Delta_m(X) = H_m(X) - \overbrace{\beta_m \odot V_m}^{\text{projected comp.}} \quad m = 1, \ldots, M \tag{8}$$

$$\Delta(X) = H(X) - \overbrace{\alpha \odot V}^{\text{projected comp.}}, \tag{9}$$

where $\alpha = [\alpha_1, \ldots, \alpha_n]$, $\beta_m = [\beta_{m,1}, \ldots, \beta_{m,n}]$, and $\odot$ denotes elementwise multiplication. Finally, the tokens in X are updated by making use of both Δ^l and Δ as given by (7).

Finally, we notice that Belief-Attention needs to introduce an additional weight matrix W^l in (7) in comparison to the standard attention. In general, the matrix W^l per attention layer would incur only a small overhead in model parameters..

Remark 1 *The recent work (Zhai, 2026) considers taking the local perpendicular component $\Delta^l(X)$ in (5) as the residual signal in an attention layer and observe noticeable performance gain in training various sizes of GPT2 for NLP. As a result, the work of (Zhai, 2026) can be considered as a special case of Belief-Attention by setting $W = 0$ in (7).*

2.2 Impact of Projected Component in a Transformer

We note from Subsection 2.1 that Belief-Attention only takes the two perpendicular components $\Delta(X)$ and $\Delta^l(X)$ as the residual signals when updating the tokens. One natural question is whether the local $\{\beta_m \odot V_m\}_{m=1}^M$ and global projected components $\alpha \odot V$ in (8)–(9) carry useful information about token correlations—specifically, whether the coefficients $\{\beta_m\}_{m=1}^M$ and α noticeably reflect these correlations.

Impact of projected component on validation performance: In order to gain insights regarding the above question, we conducted two experiments. In the first one, we performed an ablation study evaluating the impact of different residual signals in the attention layers of a Transformer on the validation accuracy when training the models over CIFAR100 for image classification.

Fig. 3 visualizes the validation accuracy over epochs for different residual signals being employed in all the attention layers of a Transformer. It is clear that both the perpendicular and projected components lead to reasonable validation performance. The results also suggest that the perpendicular component is more informative of the token-correlation than the projected components.

Fig. 3 also includes the validation performance (the curve in black color) by taking the original V vectors as the residual signals in all the attention layers of the Transformer. Unsurprisingly, the model fails to obtain any meaningful classification results regardless of the number of training epochs.

Correlation between $\angle(MH[i,:], V[i,:])$ and $\|\alpha_i V[i,:]\|$: In the second experiment, we investigate the correlation between the angle $\angle(MH[i,:], V[i,:])$ and the magnitude $\|\alpha_i V[i,:]\|$ for each token index i within an transformer block in a ViT model. In brief, the angle between the attention output $MH[i,:]$ and the original vector $V[i,:]$ reflects their directional difference.

Fig. 4 plots the relationship between the angle and magnitude within a given transformer block when a batch of samples is passed through a well-trained ViT model. Clearly, the two quantities are highly correlated: for any given angle, the magnitude $\|\alpha_i V[i,:]\|$ varies only within a narrow range. We note that this high correlation between angle and magnitude holds across all transformer blocks in the investigated ViT model.

In summary, these experimental results demonstrate that the coefficients $\{\beta_m\}_{m=1}^M$ and α strongly reflect token correlation. The key question now is how to modify Belief-Attention so that both the perpendicular and projected components can be effectively leveraged to boost performance.

3 BELIEF2-ATTENTION

In this section, we first demonstrate the benefits of introducing a two-layer FFN within Belief2-Attention to handle the projected component. We then briefly explain that the inner product QK^T can be modified to $(K+Q)K^T$ before being fed into the softmax operation to capture richer token correlations. Following this, we present and analyze the update expressions of Belief2-Attention, arguing that under certain mild conditions, it possesses higher representational capacity than standard attention.

3.1 MOTIVATION OF TWO-LAYER FFN WITHIN BELIEF2-ATTENTION

As demonstrated in Fig. 2, we propose to let the projected component $\alpha \odot V$ first go through an (element-wise) activation function $\phi(\cdot)$ and then a linear mapping W^p before merging with the considered tokens. From a high level viewpoint, a two-layer FFN is implicitly embedded within Belief2-Attention. In particular, the FFN has two linear mapping, of which the first mapping $W^V = \text{Concat}(W_1^V, \ldots, W_M^V)$ converts the tokens in X into value vectors in V (see (1)) and the second mapping W^p post-processes $\phi(\alpha \odot V)$ before merging with the considered tokens.

Intuitively, with the two-layer FFN, Belief2-Attention would have greater ability to modify tokens as desired than standard attention and Belief-Attention. As a result, when building a Transformer by stacking a set of Belief2-Attention and traditional FFN layers, the final model would have high capability to fit the input-output training data.

3.2 MOTIVATION OF $(Q+K)K^T$ BEFORE SOFTMAX OPERATION

Note that standard attention first computes $QK^\top$ and then applies a row-wise softmax operation to yield unified weighting scores. To capture richer token correlations, we propose modifying the inner product $QK^\top$ to $(Q+K)K^\top$. Intuitively, $(Q+K)K^\top$ provides more degrees of freedom for computing relative correlations: $KK^\top$ naturally models symmetric correlations, whereas $QK^\top$ captures complementary directional aspects beyond the capacity of $KK^\top$.

This design is particularly well-suited for vision tasks, where image patches are processed in a non-causal manner. The symmetric term $KK^\top$ implicitly enforces mutual, bidirectional interactions between token pairs, complementing the directional relationship captured by $QK^\top$

Remark 2 *The formulation $(Q+K)K^T$ is not the only way to model both symmetric and directional token correlations in vision tasks; alternatives such as $Q(Q+K)^T$ or $QK^T + VV^T$ could also be considered. However, $QK^T + VV^T$ incurs significant computational overhead because QK^T and VV^T need to be computed independently.*

3.3 Update expressions

Based on the earlier motivations for Belief2-Attention, its update expressions are designed to be

$$\mathrm{H}_m(X) = \mathrm{attention}(\overbrace{XW_m^Z}^{Z_m}, \overbrace{XW_m^Q}^{Q_m}, \overbrace{XW_m^K}^{K_m}, \overbrace{XW_m^V}^{V_m}) = \mathrm{softmax}\left(\frac{(Q_m + K_m)K_m^T}{\sqrt{d_m}}\right) V_m \tag{10}$$

$$\mathrm{MH}(X) = \mathrm{Concat}(\mathrm{H}_1(X), \ldots, \mathrm{H}_M(X)) \tag{11}$$

$$V(X) = \mathrm{Concat}(V_1, \ldots, V_M) \tag{12}$$

$$\Delta(X) = \mathrm{MH}(X) - \alpha \odot V(X) \tag{13}$$

$$X \Leftarrow X + \Delta(X)W + \phi(\alpha \odot V(X))W^p, \tag{14}$$

where we use $\phi(\cdot)$ to denote an (linear/nonlinear) activation function, each element α_i in α is computed by following (6). The linear mapping W^p is applied to process the projected component $\alpha \odot V(X)$ before being merged with the input token.

On relationship between Belief2-Attention and standard Attention: We now argue that Belief2-Attention has greater representational capacity than the standard Attention under certain mild conditions. Firstly, we note that (14) can be rewritten as

$$X \Leftarrow X + \overbrace{(MH(X) - \alpha \odot V(X))}^{\text{perpendicular comp.}} W + \phi(\overbrace{\alpha \odot V(X)}^{\text{projected comp.}})W^p \tag{15}$$

$$= X + (MH(X) - \alpha \odot V(X))W + \underbrace{[\phi(\alpha \odot XW^V)]W^p}_{\text{two-layar FFN}}, \tag{16}$$

where unlike that of Belief-Attention, both the perpendicular and projected components are present in the expression. No useful information is dropped on purpose in the above expression.

If we let $\phi(\cdot)$ be an identity (no nonlinear mapping is enforced) function and $W^p = W$, it is immediate that (15) reduces to the update expression for the standard Attention:

$$X \Leftarrow X + MH(X)W. \tag{17}$$

This indicates that the functional space of Belief2-Attention includes that of the standard Attention. In other words, Belief2-Attention indeed has higher representational capacity. We use $\mathcal{P}[f_\theta]$ to denote the functional space parameterized by θ. The above results can be summarized in a proposition below:

Proposition 1 *Let $\phi(\cdot)$ be an identity function. Suppose the update expression for Belief2-Attention follows from (10)-(14) while the update expression for the standard Attention is the one by setting $W^p = W$. We then have*

$$\mathcal{P}[\textit{Attention}] \subseteq \mathcal{P}[\textit{Belief2-Attention}]. \tag{18}$$

Next, we analyze the impact of W and W^p in (15) assuming $\phi(\cdot)$ is the identity function. Here, the two linear mappings act on the perpendicular and projected components independently. As a result, (15) offers more degrees of freedom for token modification than standard attention in (17). We thus expect Belief2-Attention to achieve a lower training loss than standard attention—an expectation validated by our experimental results on ImageNet classification and segmentation.

Remark 3 *In this work, we assume that the activation function $\phi(\cdot)$ in Belief2-Attention is pre-selected for specific tasks. A promising direction for future research is to learn $\phi(\cdot)$ dynamically during training rather than relying on manual selection, which could further enhance the flexibility and expressiveness of Belief2-Attention.*

Study of two-layer FFN in (16): It is worth noting that the two-layer FFN in (16) differs slightly from the conventional two-layer FFN in a Transformer, which is generally expressed as:

$$\text{FFN}(X) = \varphi(XW^1)W^2, \tag{19}$$

where $\varphi(\cdot)$ denotes an element-wise activation function, while W^1 and W^2 represent weight matrices.

Comparing (16) and (19) reveals that the FFN formulation in Belief2-Attention explicitly incorporates token correlations. Specifically, the weighting vector α preceding XW^V in (16) encodes pairwise token relationships, whereas the conventional FFN processes tokens independently without utilizing such contextual information.

3.4 Parameter and computational overhead of Belief2-Attention

As seen in (10)–(14), Belief2-Attention introduces an additional weight matrix W^p compared to standard attention. However, because Belief2-Attention implicitly incorporates a two-layer FFN, we can reduce the parameter count in the standard FFN blocks to keep the overall model size identical to that of a standard Transformer.

While Belief2-Attention incurs extra computational cost due to the orthogonal projection and its internal two-layer FFN, this overhead is partially offset by the smaller downstream FFN layers. Detailed comparisons of computational cost across different tasks are provided in Tables 2–4.

Remark 4 *In practice, a two-layer FFN can be designed by introducing a modulation mechanism into the update expression:*

$$\mathit{FFN}(X) = [(XW^3) \odot \varphi(XW^1)]W^2, \tag{20}$$

where the intermediate activations $\varphi(XW^1)$ are modulated by XW^3 prior to the second linear mapping, where W^3 is another linear layer.

Both (19) and (20) are widely employed in modern deep learning architectures (e.g., SwiGLU / Gated MLPs). Similarly, the implicit FFN in Belief2-Attention can be modified by incorporating a modulation term:

$$X \Leftarrow X + \Delta(X)W + [XW^s \odot \phi(\alpha \odot V(X))]\, W^p, \tag{21}$$

where W^s is an additional linear mapping introduced for modulation. When applying Belief2-Attention to downstream tasks, either update expression—(14) or (21)—may be selected based on performance requirements.

4 Experimental Results

We evaluated Belief2-Attention across three vision tasks: (1) image segmentation on the ADE20K dataset[1]; (2) image classification on ImageNet; and (3) image classification on CIFAR-100. All experiments were conducted on a workstation equipped with a single NVIDIA RTX A6000 GPU (48GB VRAM). Across all setups, the parameter counts of the conventional FFN layers were scaled down to keep the total model capacity comparable to the baselines.

In summary, Belief2-Attention consistently outperforms both standard attention and Belief-Attention across all three tasks. Notably, while original Belief-Attention lags behind standard attention on the segmentation task, Belief2-Attention successfully overcomes this limitation.

Finally, we conduct an extensive ablation study to isolate the contributions of $KK^\top$, conventional FFN layers, and the choice of activation function $\phi(\cdot)$. The results reveal two key insights: (a) integrating $KK^\top$ into the attention calculation significantly enhances validation accuracy; (b) removing conventional FFN layers while retaining only Belief2-Attention produces validation performance comparable to standard ViT baselines.

4.1 Image segmentation over ADE20K

The first experiment utilizes the second open-source repository listed in Table 7 to evaluate Belief2-Attention, Belief-Attention, and standard attention on image segmentation. We selected the

[1] https://github.com/CSAILVision/ADE20K

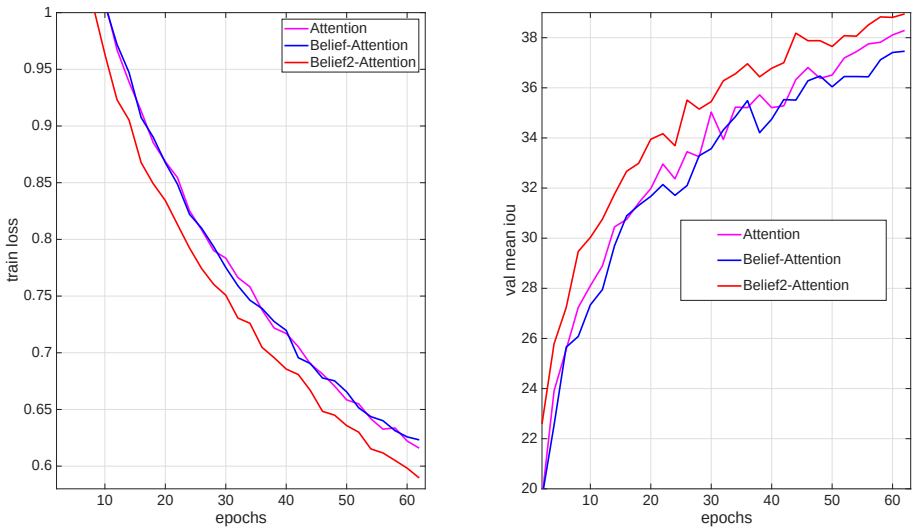


Figure 5: Performance comparison for image segmentation over the ADE20K dataset.

`vit_tiny_patch16_384` backbone to compare the performance of these three attention mechanisms. All training hyperparameters follow the default configurations of the original repository. Finally, the activation function $\phi(\cdot)$ for Belief2-Attention was set to the identity function.

Table 1: Performance comparison of three types of attention layers for image segmentation over ADE20K.

	Attention	Belief-Attention	Belief2-Attention
mIoU	38.1	37.46	**38.95**
model size	6.91M		
(training) time/iter (s)	0.31	0.33	0.32

Fig. 5 displays both the training loss and validation mIoU across epochs for performance comparison. As expected, Belief2-Attention achieves a consistently lower training loss than standard attention throughout training. This is likely because the functional space of Belief2-Attention subsumes that of standard attention, enhancing the model's capacity to fit the training data.

Furthermore, Fig. 5 and Table 1 demonstrate that Belief2-Attention yields noticeably higher validation mIoU than both Belief-Attention and standard attention. As shown in Table 1, the computational complexities of all three attention layers remain comparable.

4.2 Image Classification over ImageNet

In this experiment, we adapted the Vision Transformer backbone from the open-source JiT repository[2] for image classification on ImageNet. While the original codebase is designed for training image generation diffusion models, we selected the `JiT-B/16` variant and replaced its standard attention blocks with Belief2-Attention and Belief-Attention, respectively. The activation function $\phi(\cdot)$ for Belief2-Attention was set to the Sigmoid Linear Unit (SiLU), matching the activation used in the standard FFN layers of JiT. All models were trained from scratch with a batch size of 480 and a learning rate of 10^{-4}, using the AdamW optimizer with $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$. Following training, model performance was evaluated on the ImageNet validation set.

Table 2: Performance comparison of three types of attention layers for image classification over ImageNet.

	Attention	Belief-Attention	Belief2-Attention
val acc. (top1)	73.25	74.20	**75.02**
val acc. (top5)	91.02	91.28	**91.98**
model size	128.86M		
(training) time/iter (s)	0.358	0.362	0.360

[2] https://github.com/LTH14/JiT

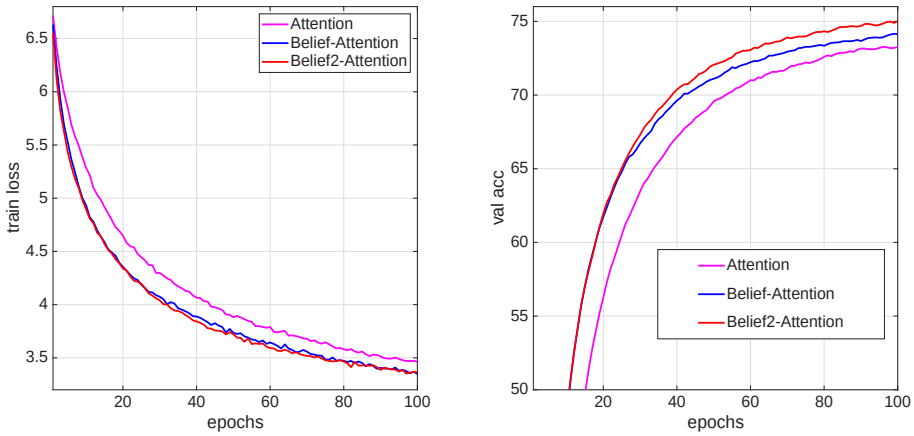


Figure 6: Performance comparison for image classification over ImageNet of 1000 classes.

Table 2 reports the validation accuracy, parameter count, and training overhead for all three attention variants. The results demonstrate that, given a comparable parameter budget, Belief2-Attention yields superior validation accuracy at the cost of a minor computational overhead.

Fig. 6 visualizes both the training loss and validation accuracy across epochs. Notably, both Belief2-Attention and Belief-Attention outperform standard attention considerably as training progresses. The faster convergence of the Belief2-Attention model can be attributed to the implicit two-layer FFN embedded within each attention layer, which enhances the model's representational capacity.

4.3 Image classification over CIFAR100

In this experiment, we adopted the first open-source repository listed in Table 7 to train a ViT model on CIFAR-100. We replaced the standard attention layers with Belief-Attention and Belief2-Attention for performance comparison. The activation function $\phi(\cdot)$ for Belief2-Attention was set to the Gaussian Error Linear Unit (GELU), matching the activation used in the conventional FFN layers of the ViT model.

Table 3: Performance comparison of three different attention layers for image classification over CIFAR100.

	Attention	Belief-Attention	Belief2-Attention
val acc.	72.36±0.13	74.18±0.25	**75.90**±0.50
model size		3.04M	
time/epoch (s)	43.24	52.84	50.58

Aside from the model modifications mentioned above, all training setups followed the original open-source implementation. Specifically, each model was trained for 100 epochs using the AdamW optimizer. To account for random variation, three independent training runs were performed for each setup across random seeds $\{0, 50, 100\}$.

Table 3 reports the validation accuracy, parameter count, and training overhead for each model. The results show that Belief2-Attention yields significantly higher validation accuracy than both standard attention and Belief-Attention, at the cost of a slight computational overhead relative to standard attention. This confirms that embedding a two-layer FFN directly within the attention mechanism effectively boosts model performance.

4.4 Ablation study of Belief2-Attention

In this subsection, we conduct an extensive ablation study to evaluate the core design choices of Belief2-Attention on CIFAR-100 and ImageNet. Specifically, we investigate the impact of the activation function $\phi(\cdot)$, the $KK^\top$ term, and conventional FFN layers on overall classification performance.

Impact of $\phi(\cdot)$: We evaluated three activation functions for $\phi(\cdot)$ in Belief2-Attention when training ViT models on CIFAR100: GELU, SiLU, and the identity function. Table 4 summarizes the experimental results. As shown in the table, the two non-linear activations, GELU and SiLU, produce considerably higher validation accuracy than the identity function. This confirms that non-linear

activation functions indeed increase the representational capacity of the model. Finally, comparing Tables 3 and 4, we note that Belief2-Attention using the identity function still outperforms standard attention by a substantial margin.

Table 4: Impact of different activation functions for $\phi(\cdot)$ in Belief2-Attention based ViT over CIFAR100. $(Q+K)K^T$ was implemented in Belief2-Attention.

	GELU(·)	SiLU(·)	identity fun.
val acc.	75.90±0.50	75.87± 0.14	74.39±0.40
time/epoch (s)	50.58	50.38	49.86

Impact of KK^T and traditional FFN layers: We also investigated the performance of Belief2-Attention under two specialized configurations: (a) omitting the $KK^\top$ matrix, and (b) removing conventional FFN layers entirely. Because Belief2-Attention already incorporates an internal FFN, it is of particular interest to evaluate the capacity of a model constructed solely by stacking Belief2-Attention layers.

Tables 5 and 6 summarize the performance of the Belief2-Attention-based ViT model under different configurations on CIFAR-100 and ImageNet, respectively. First, the results in both tables demonstrate that incorporating $KK^\top$ into the attention calculation of Belief2-Attention boosts validation performance significantly. This can be attributed to $KK^\top$ capturing symmetric token correlations that are beyond the expressive capacity of $QK^\top$.

Second, comparing Tables 2, 3, 5, and 6 reveals that even when completely removing conventional FFN layers or omitting $KK^\top$, model performance remains superior to that of the standard Transformer.

These findings demonstrate that a model built exclusively by stacking Belief2-Attention layers achieves robust performance. This pure Belief2-Attention architecture offers an attractive alternative to standard Vision Transformers when compact model size and parameter constraints are prioritized.

Table 5: Ablation study of different Belief2-Attention-based ViT configurations on CIFAR-100. The notation "w.o. FFN" indicates that all conventional FFN layers are removed from the ViT architecture, leaving only the embedded FFNs within Belief2-Attention.

	w. $(Q+K)K^T$ w. FFN	w. $(Q+K)K^T$ w.o. FFN	w. QK^T w. FFN
val acc.	75.90±0.50	73.79± 0.10	73.81±0.24
model size	3.04M	1.71M	3.04M
time/epoch (s)	50.58	40.94	50.37

Table 6: Ablation study for different setups of Belief2-Attention based ViT over ImageNet.

	w. $(Q+K)K^T$ w. FFN	w. $(Q+K)K^T$ w.o. FFN	w. QK^T w. FFN
val acc. (top1)	75.02	73.59	73.73
time/iter (s)	0.360	0.278	0.357

5 Conclusions

In this work, we first demonstrate via an ablation study that the projected component from the orthogonal projection in Belief-Attention carries critical token correlation information that should not be discarded. We then propose Belief2-Attention, which fully utilizes both the perpendicular and projected components by seamlessly integrating a two-layer FFN inside the attention layer to process the projected component. We show that under mild conditions, Belief2-Attention possesses greater representational capacity than standard attention. Because Belief2-Attention embeds an internal FFN, the total parameter count can be preserved by reducing the size of conventional FFN layers. Additionally, we modify the inner-product term from $QK^\top$ to $(Q+K)K^\top$, enabling $KK^\top$ and $QK^\top$ to model symmetric and directional correlations, respectively. Experimental results on image segmentation and classification demonstrate that Belief2-Attention consistently outperforms standard attention, whereas Belief-Attention struggles on image segmentation. Finally, an ablation study indicates that models constructed solely with Belief2-Attention layers achieve performance comparable to standard ViT baselines.

AI USE STATEMENT

Generative AI tools were used solely to assist with proofreading, language refinement, and draft feedback. All scientific claims, experimental results, analyses, and final text were thoroughly reviewed by the authors, who take full responsibility for the content of this submission.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Florencia Leoni Aleman Lama Ahmad, Ilge Akkaya, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Lenny Bogdonoff Christopher Berner, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, and Shawn Jain. Gpt-4 technical report. arXiv:2307.09288 [cs.CL], 2023.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. arXiv:2004.05150v2, 2020.

T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691, 2023.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborna, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly abd J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

A. Hatamizadeh, J. Song, G. Liu, J. Kautz, and A. Vahdat. DiffiT: Diffusion Vision Transformers for Image Generation. In *ECCV*, 2024.

N. Kitaev, A. Kaiser, and A. Levskaya. Reformer: The Efficient Transformer. In *ICLR*, 2020.

S. Latif, A. Zaidi, H. Cuayahuitl AZ, F. Shamshad, M. Shoukat, and J. Qadir. Transformers in speech processing: A survey. arXiv:2303.11607 [cs.CL], 2023.

H. Liu, M. Zaharia, and P. Abbeel. Ring attention with blockwise transformers for near-infinite context. arXiv:1706.03762 [cs. CL], 2023.

W. Peebles and S. Xie. Scalable Diffusion Models with Transformers. In *ICCV*, 2023.

A. Sun, W. Zhao, X. Han, C. Yang, Z. Liu, C. Shi, and M. Sun. Burstattention: An efficient distributed attention framework for extremely long sequences. arXiv:2403.09347 [cs.DC], 2023.

H. Touvron, L. Martin, P. Albert K. Stone, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Yinghai Lu Diana Liskovich, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Andrew Poulton Yixin Nie, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Jian Xiang Kuan Adina Williams, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Aurelien Rodriguez

Sharan Narang, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288 [cs.CL], 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. arXiv:1706.03762 [cs. CL], 2017.

S. Wang, M. Khabsa B. Z. Li, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. arXiv:2006.04768v3, 2020.

Shuangfei Zhai. Exclusive self attention. arXiv:2603.09078 [cs.LG], 2026.

Guoqiang Zhang. Beliefformer: Belief-attention in transformer. https://openreview.net/forum?id=Ard2QzPAUK, 2025.

Guoqiang Zhang, J. P. Lewis, and W. Bastiaan Kleijn. Exact diffusion inversion via bidirectional integration approximation. arXiv:2307.10829 [cs.CV], 2023.

APPENDICES

CIFAR10 & CIFAR100	https://github.com/aanna0701/SPT_LSA_ViT
Image segmentation	https://github.com/rstrudel/segmenter

Table 7: list of open-source repositories expoited in this paper.

A SCATTER PLOTS OF THE ANGLE $\angle(MH[i,:], V[i,:])$ VERSUS $\|\alpha_i V[i,:]\|$

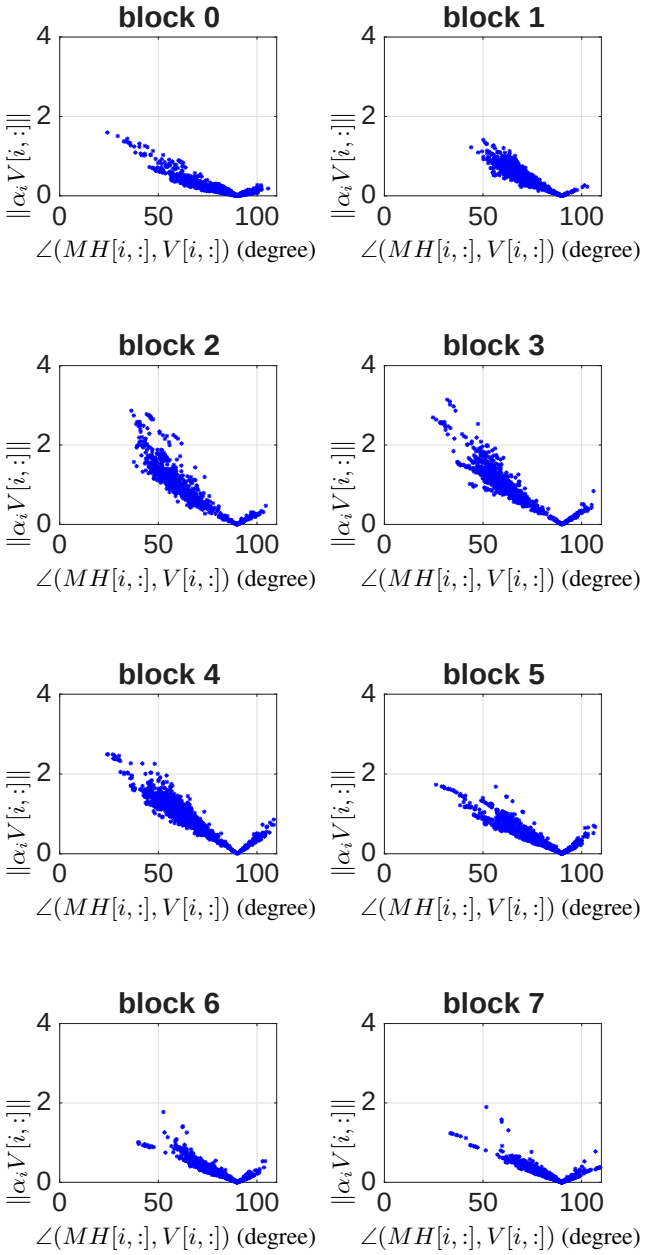


Figure 7: Scatter plots of the angle$\angle(MH[i,:], V[i,:])$ versus $\|\alpha_i V[i,:]\|$ for each token index i across the first eight transformer blocks after training the standard ViT model.