RCViT: Encoding Spatial Structure through Alternating Row–Column Attention

Anonymous authors
Paper under double-blind review

Abstract

Positional encodings provide Vision Transformers with information about the spatial arrangement of image patches. Spatial relationships can also be specified through the attention pattern itself: row and column connections represent horizontal and vertical alignment, and alternating these connections across layers composes dependencies along both image axes. We investigate this observation through RCViT, a Vision Transformer with sparse row–column attention. Deterministic masks combine local and dilated neighbors, alternate the active axis across layers, and assign directional roles to attention heads. Global heads preserve unrestricted interactions, controlling the strength of the spatial inductive bias. The construction adds no trainable parameters and integrates with absolute, relative, conditional, and rotary positional encodings. Classification experiments demonstrate the utility of the attention pattern without positional embeddings and its complementarity with positional encoding in mixed directional–global models. Ablations and resolution and layout evaluations identify the importance of axis alternation and the dependence on head allocation and model configuration. The results support alternating row–column attention as a spatial inductive bias complementary to global attention and positional encoding.

1 Introduction

Vision Transformers (ViTs) model interactions between image patches through self-attention (Dosovitskiy et al., 2021; Vaswani et al., 2017). Without positional information, unrestricted self-attention is permutation equivariant: it processes patch tokens without distinguishing their spatial arrangement. Positional embeddings, relative position biases, and rotary embeddings incorporate position into token features or attention scores (Wu et al., 2021b; Chu et al., 2023; Heo et al., 2024). Attention patterns complement these representations by specifying which patches can interact.

Our starting point is that horizontal and vertical alignment are themselves spatial relationships. Patches on the same row share one spatial coordinate, while patches on the same column share the other. Row and column attention therefore impose structure on feature aggregation even when no positional embedding is added. Alternating the axes across successive layers composes these one-dimensional interactions into two-dimensional dependencies. *The spatial information lies in the arrangement of the connections, not only in an explicit representation of position.* This observation motivates studying axis alternation as a spatial inductive bias in a plain ViT.

Spatially structured attention has important precedents. Axial attention factorizes interactions along individual dimensions (Ho et al., 2019; Wang et al., 2020), CSWin uses horizontal and vertical stripes (Dong et al., 2022), and LookHere combines directed attention with unrestricted heads (Fuller et al., 2024). We study sparse row–column neighborhoods and the strength of their constraint. Restricting every head to an axis introduces a strong spatial prior, but also removes direct interactions outside those neighborhoods. This raises the question of how directional structure can complement global attention when positional information is already available.

We introduce RCViT to study this question. Directional heads attend to local and dilated neighbors along one image axis, with rows and columns alternating across blocks. Deterministic masks specify these neighborhoods, and a cyclic assignment distributes directional roles across heads. The remaining heads use global attention, retaining direct interactions between arbitrary patches. Varying the number of directional heads controls the spatial constraint without changing the learned projections or adding

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.

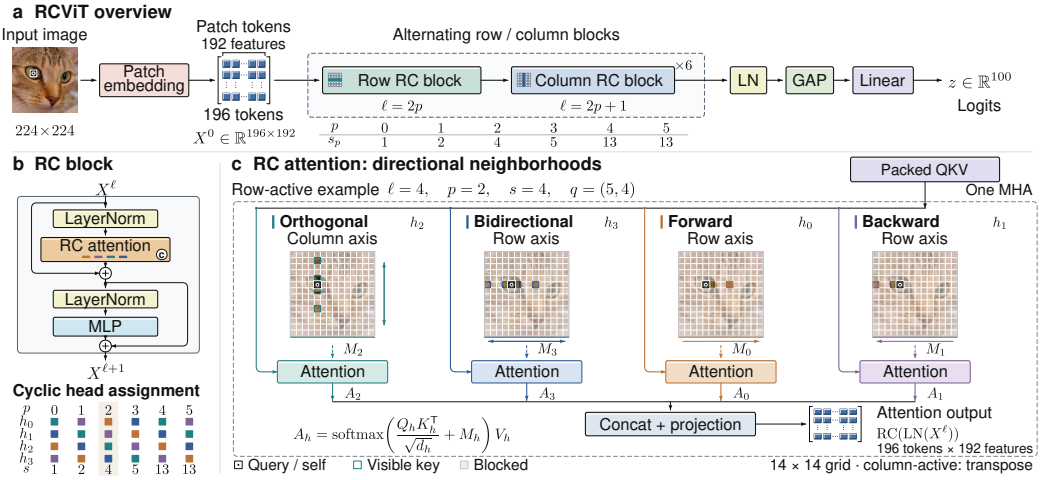

Figure 1: **RCViT architecture and directional attention.** (a) Illustrative ImageNet-100 configuration ($D = 192$, $H = 4$): blocks alternate row and column attention; matrices contain 196 token rows and 192 feature columns. (b) Standard pre-normalization residual block. (c) Fully directional attention uses orthogonal-axis, bidirectional, forward, and backward neighborhoods, with cyclic head assignment across layer pairs. Colored patches indicate admissible keys, not attention weights. Mixed attention retains k directional and $H - k$ global heads; $k = 1$ retains only the backward neighborhood.

trainable parameters. Because the masks restrict connections rather than replace positional operations, they integrate with absolute, relative, conditional, and rotary positional encodings.

The experiments connect this construction to its intended spatial role. Without positional embeddings, row–column attention improves recognition on three small-image datasets. With positional encoding, a single directional head improves all four evaluated ImageNet-100 baselines, highlighting the value of retaining global interactions alongside the spatial constraint. ImageNet-1K comparisons further characterize how this complementarity depends on the positional encoding and backbone. Attention-pattern ablations isolate axis alternation from generic sparsity and head assignment; resolution and layout evaluations examine generalization beyond the training geometry. Together, the results support combining an explicit spatial constraint with unrestricted interactions.

Our contributions are as follows:

- We introduce RCViT, which encodes horizontal and vertical relationships through alternating sparse row–column attention, with global heads controlling the spatial constraint.
- We formulate the construction as parameter-free attention masks compatible with absolute, relative, conditional, and rotary positional encodings in a standard Transformer block.
- We characterize the role of spatial constraints through classification, head-allocation and attention-pattern ablations, and generalization to changes in image resolution and layout.

2 Related Work

Spatial inductive biases in self-attention. Permutation-equivariant attention is well suited to learning on sets (Lee et al., 2019), whereas visual recognition also benefits from image-specific structure. Connections between self-attention and convolution have been studied in terms of expressive power (Cordonnier et al., 2020) and local attention operators (Ramachandran et al., 2019). ConViT uses gated positional self-attention to introduce a soft convolutional inductive bias (D'Ascoli et al., 2021); CvT incorporates convolutions into token embeddings and projections (Wu et al., 2021a). RCViT introduces a spatial constraint through fixed attention neighborhoods. Global heads complement this constraint without requiring a learned gate or an additional positional module.

Positional encoding in vision Transformers. ViT adds learned absolute position embeddings to patch tokens (Dosovitskiy et al., 2021). Relative position encoding represents the displacement between a query and a key, often through the attention logits (Shaw et al., 2018; Wu et al., 2021b; Liu et al., 2021). CPVT generates conditional positional encodings from local features (Chu et al., 2023). Rotary position embedding (RoPE) applies position-dependent rotations to queries and keys (Su et al., 2024); RoPE-ViT adapts this operation to visual recognition (Heo et al., 2024), and nD-RoPE extends it to multidimensional positions (Li et al., 2026). These mechanisms modify features or attention scores. RCViT instead restricts the available connections, allowing the spatial constraint and the positional mechanism to operate within the same head.

Axial and spatially restricted attention. Axial attention decomposes multidimensional interactions along individual axes (Ho et al., 2019), and Axial-DeepLab develops position-sensitive axial attention for dense prediction (Wang et al., 2020). CCNet aggregates row and column context through recurrent criss-cross attention (Huang et al., 2019); CSWin assigns horizontal and vertical stripes to groups of heads (Dong et al., 2022). RCViT builds on the same geometric principle but uses sparse local and dilated offsets, rather than complete axes or stripes, and alternates the active axis.

Other designs combine locality with wider context through shifted windows (Liu et al., 2021), blocked local attention (Vaswani et al., 2021), sliding-window and dilated neighborhoods (Hassani et al., 2023; Hassani & Shi, 2022), or locally grouped and global sub-sampled attention (Chu et al., 2021). Our study concerns the allocation of sparse axis-specific and unrestricted heads within a plain ViT block, holding the backbone and positional mechanism fixed.

Directed attention and global interactions. LookHere introduces orientation and distance information through fixed angular fields of view and distance-dependent attention biases (Fuller et al., 2024). Like LookHere, RCViT combines directional and unrestricted attention. Its directional neighborhoods are sparse and one-dimensional, with alternating row and column axes and cyclic head assignments, rather than angular regions with distance penalties. We study this construction both without positional embeddings and in combination with existing positional mechanisms.

3 METHOD

3.1 SPATIAL INFORMATION IN THE ATTENTION PATTERN

Let an image be partitioned into a grid $\mathcal{G} = \{0, \dots, R-1\} \times \{0, \dots, C-1\}$, with $N = RC$ patch tokens and feature matrix $X \in \mathbb{R}^{N \times D}$. For one attention head, a mask $M \in \{0, -\infty\}^{N \times N}$ defines the admissible query–key pairs:

$$\mathcal{A}(X; M) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + M\right) V, \qquad Q = XW_Q,\ K = XW_K,\ V = XW_V. \tag{1}$$

Here d_h is the head dimension, and softmax is applied to each query row. Unrestricted attention corresponds to $M = 0$.

For a token permutation matrix P, attention without positional encoding satisfies

$$\mathcal{A}(PX; PMP^\top) = P\mathcal{A}(X; M). \tag{2}$$

When $M = 0$, permuting the input only permutes the output. A fixed row–column mask preserves this equivariance for permutations that preserve its spatial connections, but not for arbitrary permutations in general. It therefore introduces a spatial inductive bias without adding positional embeddings to X. The mask encodes relationships between patches, rather than a unique absolute coordinate for each token; Appendix A gives the derivation and its scope.

3.2 ALTERNATING ROW AND COLUMN NEIGHBORHOODS

Layer indices start at $\ell = 0$. The active axis and layer-pair index are

$$a_\ell = \begin{cases} \text{row}, & \ell \text{ even}, \\ \text{column}, & \ell \text{ odd}, \end{cases} \qquad p_\ell = \lfloor \ell/2 \rfloor \bmod 6. \tag{3}$$

Let $e_{\text{row}} = (0, 1)$ and $e_{\text{column}} = (1, 0)$. For an axis of length $A \geq 2$, a six-entry schedule determines the dilation $s_p(A)$. At the 14×14 training grid, the schedule is $(1, 2, 4, 5, 13, 13)$; its extension to other grid sizes is specified in Appendix A. Both layers in a pair share the schedule index.

For dilation s, the bidirectional, forward, and backward offset sets are

$$\Delta_{\leftrightarrow}(s) = \{-s, -1, 0, 1, s\}, \quad \Delta_{+}(s) = \{0, 1, s\}, \quad \Delta_{-}(s) = \{-s, -1, 0\}. \tag{4}$$

Given a query at $u = (r, c)$, axis a, and offset set Δ, its neighborhood is

$$\mathcal{N}_{a,\Delta}(u) = \{u + \delta e_a : \delta \in \Delta\} \cap \mathcal{G}. \tag{5}$$

Self-attention is always included. Off-grid locations are discarded, with no wraparound. Duplicate offsets, such as those at $s = 1$, denote the same key. Immediate neighbors preserve local interactions, while dilated neighbors extend the spatial range without increasing the number of admissible offsets.

Alternating the axes allows dependencies to extend in both spatial dimensions. For example, when the corresponding offsets are admissible, two consecutive layers permit a dependency along

$$(r, c) \longrightarrow (r, c + \delta_c) \longrightarrow (r + \delta_r, c + \delta_c). \tag{6}$$

This composition is the spatial role of alternation. Since the neighborhoods are sparse, it does not establish complete global connectivity after only two layers.

3.3 Directional roles and global attention heads

Let H denote the number of heads and $k \in \{0, \ldots, H\}$ the number of directional heads. The other $H - k$ heads use unrestricted attention. The fully directional four-head model assigns orthogonal, bidirectional, forward, and backward neighborhoods cyclically across layer pairs (Figure 1; Appendix A.1). Mixed attention makes the first $H - k$ heads in this ordering unrestricted. The cyclic assignment changes head roles without moving or tying projection parameters. For $k = 1$ and any evaluated $H \geq 3$, the remaining directional head uses backward offsets on the active axis:

$$h_\ell^* = (p_\ell + H - 1) \bmod H, \qquad \mathcal{N}_{\ell h}(u) = \begin{cases} \mathcal{N}_{a_\ell, \Delta_-(s)}(u), & h = h_\ell^*, \\ \mathcal{G}, & h \neq h_\ell^*. \end{cases} \tag{7}$$

This definition applies to the three- and four-head ViT-T and six-head ViT-S models in the ImageNet-1K study. The directional head remains axis-dependent; global heads allow interactions between any pair of patches. The four-head allocation study evaluates every k, including fully global attention ($k = 0$) and fully directional attention ($k = 4$).

3.4 Compatibility with positional encodings

Let ϕ denote a positional mechanism, $Q_{\ell h}^{\phi}$, $K_{\ell h}^{\phi}$, and $V_{\ell h}^{\phi}$ its resulting queries, keys, and values, and $B_{\ell h}^{\phi}$ any additive positional bias. The head output is

$$O_{\ell h} = \operatorname{softmax}\left(\frac{Q_{\ell h}^{\phi} (K_{\ell h}^{\phi})^{\top}}{\sqrt{d_h}} + B_{\ell h}^{\phi} + M_{\ell h} \right) V_{\ell h}^{\phi}, \tag{8}$$

where $[M_{\ell h}]_{uv} = 0$ if $v \in \mathcal{N}_{\ell h}(u)$ and $-\infty$ otherwise. Absolute position embeddings and conditional encodings modify the token features used to compute queries, keys, and values. Relative bias contributes B^{ϕ}, whereas rotary embeddings transform queries and keys while leaving the value projection unchanged. RCViT preserves the corresponding operations and changes only M.

The two operations constrain attention differently. Positional encoding influences the scores of admissible keys, whereas the mask excludes connections independently of their content scores. Attention weights remain content-dependent within each neighborhood. The output projection combines directional features with features from global heads, which retain access to off-axis and distant patches. Thus, alternating row–column attention specifies spatial relationships without replacing either content-dependent aggregation or the positional mechanism.

The model retains standard multi-head projections, pre-normalization, residual connections, and feed-forward layers. Classification uses layer normalization, global average pooling, and a linear head. No auxiliary objective or learned mask is introduced, so parameter counts are unchanged. The current implementation evaluates the masks with dense attention; exploiting their sparsity for computational savings would require a sparse implementation.

Table 1: **Directional attention improves recognition without positional embeddings.** Best-EMA held-out validation top-1 accuracy (%). Every model omits positional embeddings. Values are mean ± sample standard deviation over five runs for CIFAR-10 and SVHN and three for CIFAR-100. The complete allocation study appears in Appendix C.

Dataset	Global ($k = 0$)	RCViT ($k = 4$)	RCViT ($k = 1$)
CIFAR-10	77.50±0.26	89.58±0.11	88.24±0.37
SVHN	67.00±1.95	95.73±0.09	95.73±0.12
CIFAR-100	50.25±0.34	66.43±0.36	61.21±0.57

4 EXPERIMENTS

Our experiments address three questions: whether alternating row–column attention is useful without positional embeddings, whether it complements positional encoding, and how its spatial constraint affects recognition and generalization under changes in input geometry.

4.1 EXPERIMENTAL SETUP

Datasets and models. We first examine the spatial inductive bias on CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and SVHN (Netzer et al., 2011) without positional embeddings. These experiments use a fixed 10% holdout of the training split. We then evaluate classification on the 100-class ImageNet subset used in contrastive multiview coding (Tian et al., 2020) and on ImageNet-1K (Deng et al., 2009). The ImageNet-100 validation set contains 5,000 images; ImageNet-1K uses the standard 50,000-image validation set. All classifiers use 16×16 patches and 224×224 training images. The small-image and ImageNet-100 model has 12 blocks, width 192, and four heads. ImageNet-1K uses ViT-T with width 192 and three or four heads, and ViT-S with width 384 and six heads. Both have 12 blocks. All models use global average pooling.

Training and reporting. Within each comparison, we match the backbone, positional mechanism, data split, augmentation, optimizer, and training duration. Small-image models train for 120 epochs, and their metric is the highest EMA accuracy on the training holdout. ImageNet-100 and ImageNet-1K models train for 400 and 300 epochs, respectively, and use the final exponential moving average (EMA) model. ImageNet training uses AdamW (Loshchilov & Hutter, 2019), a peak learning rate of 5×10^{-4}, weight decay 0.05, five warmup epochs, and cosine decay. The effective batch size is 512. Additional details are provided in Appendix B. Tables report means and sample standard deviations across independent training runs: five for CIFAR-10 and SVHN, and three for CIFAR-100 and ImageNet unless otherwise indicated. Baseline–RCViT[1] pairs use matched random initializations; We compare positional mechanisms within a common implementation, rather than published results under different training recipes.

Comparisons. For each positional mechanism, the $k = 0$ model is the baseline and the $k > 0$ model adds the RCViT masks. We evaluate learned absolute position embeddings (APE), learned relative position bias, CPVT conditional positional encoding (CPVT-CPE), and nD-RoPE. On ImageNet-100, we examine fully directional attention and the full head allocation curve. The $k = 1$ configuration was selected on ImageNet-100 with nD-RoPE, then applied to the remaining positional mechanisms and the ImageNet-1K comparisons.

4.2 SPATIAL INDUCTIVE BIAS WITHOUT POSITIONAL EMBEDDINGS

Removing positional embeddings isolates the contribution of the attention pattern. Under this setting, fully directional attention outperforms unrestricted attention on all three datasets (Table 1). The improvement is consistent with the motivation in Section 3: restricting connections according to row and column relationships gives feature aggregation access to image geometry, even without adding position vectors to the patch tokens.

[1]Source code is available in https://anonymous.4open.science/r/RCViT/.

Table 2: **One directional head complements four positional mechanisms on ImageNet-100.** Final-EMA top-1 accuracy (%), mean ± sample standard deviation over three runs. All models have four heads. Global denotes $k = 0$, fully directional denotes $k = 4$, and the proposed mixed configuration uses $k = 1$. Δ is the mean change from the global baseline in percentage points.

Position	Global	Directional ($k = 4$)	RCViT ($k = 1$)	Δ
Learned APE	82.41 ± 0.16	81.55 ± 0.31	$\mathbf{84.79 \pm 0.68}$	+2.38
Relative bias	83.55 ± 0.35	82.11 ± 0.31	$\mathbf{84.67 \pm 0.42}$	+1.12
CPVT-CPE	83.73 ± 0.17	82.10 ± 0.09	$\mathbf{85.36 \pm 0.23}$	+1.63
nD-RoPE	85.59 ± 0.17	82.03 ± 0.40	$\mathbf{86.03 \pm 0.24}$	+0.43

Table 3: **ImageNet-1K classification with matched backbones.** Final-EMA top-1 accuracy (%) after 300 epochs. Values are mean ± sample standard deviation over three runs; RCViT constrains one head and leaves the others global. Δ denotes RCViT minus baseline.

Backbone	Heads	Position	Global	+ RCViT	Δ (pp)
ViT-T/16	3	Learned APE	70.13	70.82	+0.68
ViT-T/16	3	nD-RoPE	72.09	71.62	-0.47
ViT-T/16	4	nD-RoPE	72.56	72.46	-0.10
ViT-S/16	6	nD-RoPE	79.54	80.10	+0.56

The preferred constraint strength differs across datasets. CIFAR-10 and CIFAR-100 favor four directional heads, whereas SVHN attains its highest mean accuracy with two directional and two global heads (Appendix C). We next examine this balance with positional encoding and natural-image classification at larger scale.

4.3 Complementarity with positional encodings

Positional encoding and structured connections provide complementary sources of spatial information. On ImageNet-100, adding one directional head improves all four positional baselines, with mean gains of 0.43–2.38 percentage points (Table 2). The mixed configuration outperforms both global and fully directional attention in each of the three paired runs. The global baselines also exceed the fully directional models, indicating that retaining unrestricted interactions is important in this setting.

The comparison supports combining directional neighborhoods with global context: one head aggregates row–column features, while global heads retain direct access to the rest of the image.

The small-image experiments provide a complementary setting (Appendix C.2). Fully directional attention improves all four positional baselines on both CIFAR datasets; on SVHN, gains vary and nD-RoPE is essentially unchanged. The contrast with ImageNet-100 indicates that positional encoding alone does not determine the preferred constraint strength. Head allocation should instead be considered jointly with the classification setting and positional mechanism.

4.4 ImageNet-1K classification

ImageNet-1K tests the same construction under a larger training set and label space (Table 3). With learned APE, one directional head improves ViT-T by 0.68 percentage points. With nD-RoPE, ViT-S improves from 79.54% to 80.10%, while the global baseline retains an accuracy advantage of 0.47 and 0.10 points in the three- and four-head ViT-T configurations, respectively. Each pair has the same parameter count and uses the same 300-epoch training recipe.

The nD-RoPE results motivate studying head allocation jointly with backbone capacity. ViT-T and ViT-S exhibit different responses to the same one-directional-head construction; however, width and head count vary together, so these comparisons do not isolate their individual contributions.

Table 4: **Global-head allocation on ImageNet-100.** Final-EMA top-1 accuracy (%), mean ± sample standard deviation over three runs. The total number of heads is four, so $k = 4 - g$ for g global heads.

Global heads g	Learned APE	Relative bias	CPVT-CPE	nD-RoPE
0	81.55±0.31	82.11±0.31	82.10±0.09	82.03±0.40
1	84.75±0.40	84.85±0.37	85.06±0.27	85.42±0.15
2	84.97±0.09	84.98±0.52	85.32±0.33	85.57±0.29
3	84.79±0.68	84.67±0.42	85.36±0.23	86.03±0.24
4	82.41±0.16	83.55±0.35	83.73±0.17	85.59±0.17

Table 5: **Attention-pattern ablations with nD-RoPE.** ImageNet-100 final-EMA top-1 accuracy (%), mean ± sample standard deviation over three runs. Except for the global reference, each model has one constrained head and three global heads.

Attention pattern	Global heads	Top-1 (%)
Global attention	4	85.59±0.17
Random neighborhood	3	85.49±0.20
Local neighborhood	3	85.83±0.17
Row attention only	3	85.73±0.25
Column attention only	3	85.42±0.11
Fixed directional head	3	**86.13±0.34**
RCViT (alternating rows and columns)	3	86.03±0.24

4.5 Effect of the number of global heads

Table 4 examines the balance between spatial constraints and global interactions on ImageNet-100. Mixed configurations consistently outperform fully directional attention. Learned APE and relative bias attain their highest means with two global heads, while CPVT-CPE and nD-RoPE favor three; the two strongest CPVT-CPE allocations differ by only 0.04 points. One directional head is therefore an effective common configuration, although the preferred allocation varies. The allocation is fixed during training and inference.

The allocation curve also separates the benefit of retaining global context from that of adding a directional constraint. Restoring one global head improves every fully directional model, while retaining one directional head improves every fully global baseline. Neither endpoint alone captures the strongest observed balance.

4.6 Alternative attention patterns

To distinguish axis alternation from a generic attention restriction, we compare the $k = 1$ nD-RoPE model with alternative masks on ImageNet-100 (Table 5). The controls use random, local, row-only, or column-only neighborhoods, or preserve axis alternation with a fixed directional head. Random and local masks match the number of admissible keys for each query and preserve self-connections. All constrained models retain three global heads and use the same training recipe.

Axis alternation provides higher mean accuracy than either axis alone: 86.03%, compared with 85.73% for row-only and 85.42% for column-only attention. It also exceeds the local and random controls, suggesting that the arrangement of the connections matters beyond restricting attention to a small neighborhood. Fixing the directional head index while preserving axis alternation gives $86.13 \pm 0.34\%$, compared with $86.03 \pm 0.24\%$ for cyclic assignment. The improvement over global attention under both assignments supports axis alternation as their shared source of spatial structure.

4.7 Resolution and layout generalization

We evaluate the ImageNet-100 classifiers without fine-tuning at nine square resolutions from 160 to 512 pixels, four rectangular resolutions, and nine padded image placements. Padding places an unchanged 224×224 image on a larger canvas, separating its content scale from the surrounding

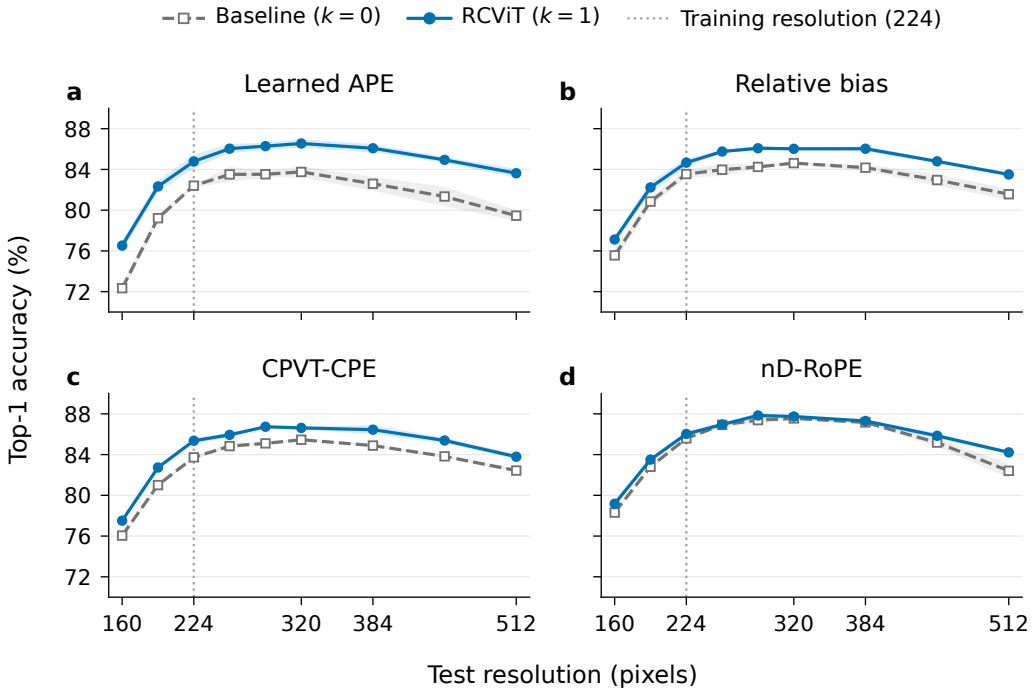

Figure 2: **Resolution generalization on ImageNet-100 without fine-tuning.** Curves show mean final-EMA top-1 accuracy over three runs; shaded bands indicate sample standard deviation. Each panel compares the global baseline with one directional head under the same positional mechanism. The vertical line marks the training resolution of 224 pixels.

Table 6: **Generalization under changes in resolution and layout.** ImageNet-100 top-1 accuracy (%), averaged within each transformation group for each run and then summarized as mean ± sample standard deviation across three runs. Resolution averages eight non-training square sizes; aspect ratio averages four rectangular sizes; canvas averages nine padded placements.

Position	Attention	Resolution	Aspect ratio	Canvas
Learned APE	Global	80.72±0.48	82.67±0.23	78.39±0.35
	+ RCViT	84.05±0.41	85.36±0.44	81.93±0.33
Relative bias	Global	82.24±0.40	83.59±0.36	79.45±0.68
	+ RCViT	83.94±0.20	85.16±0.10	80.91±0.35
CPVT-CPE	Global	82.95±0.06	83.95±0.11	81.04±0.31
	+ RCViT	84.40±0.14	85.42±0.18	82.50±0.42
nD-RoPE	Global	84.72±0.17	86.42±0.17	83.27±0.42
	+ RCViT	85.32±0.07	86.52±0.11	83.79±0.15

canvas dimensions. Appendix E specifies the complete evaluation, including translation, transpose, and rotation tests. We report accuracy and its change relative to the 224×224 evaluation to distinguish a higher clean baseline from a smaller performance loss under transformation.

RCViT has higher mean accuracy in all three transformation groups for each positional mechanism; learned-APE canvas accuracy increases from 78.39% to 81.93%. Mean accuracy is also higher at every tested square resolution (Figure 2). Relative to clean accuracy, CPVT-CPE shows a slightly larger drop in each group, as does nD-RoPE under aspect-ratio changes. Appendix E reports these paired comparisons and orientation tests. The results support improved accuracy under the evaluated geometries, rather than general spatial invariance. Paired examples in Appendix E.2 illustrate image-level responses to canvas changes without altering image content.

Table 7: **Parameter count and measured optimizer-step cost.** Baseline and RCViT have identical parameter counts. Step-time ratios are RCViT/global, reported as mean ± sample standard deviation over five paired trials. Timings include forward and backward computation and optimizer and EMA updates, but exclude data loading.

Position	Params (M)	Added params	Step time ratio
Learned APE	5.543	0	0.901 ± 0.135
Relative bias	5.509	0	1.044 ± 0.086
CPVT-CPE	5.508	0	0.907 ± 0.132
nD-RoPE	5.506	0	1.043 ± 0.027

4.8 Parameters and computational cost

RCViT preserves the parameter count of each positional baseline. We measure synchronized optimizer steps on an NVIDIA RTX 5880 Ada Generation GPU using five paired trials and GPU-resident inputs to exclude data loading (Appendix B).

Mean step-time ratios range from 0.901 to 1.044 (Table 7), indicating comparable optimizer-step cost under this protocol, with variation across trials. RCViT adds no parameters, and mean step-time increases remain small where the ratio exceeds one. These measurements do not imply end-to-end or sparse-kernel acceleration.

5 Discussion

The results distinguish the presence of a spatial prior from its appropriate strength. Alternating row and column neighborhoods introduces image geometry into feature aggregation, while global heads retain unrestricted dependencies. The ImageNet-100 allocation study supports combining these operations. Variation across ImageNet-1K configurations further motivates adapting their balance to the backbone and positional encoding.

Axis alternation composes spatial relationships across layers, whereas head assignment determines which projections operate on them. Improvements under both fixed and cyclic assignments support the connectivity pattern as their shared spatial inductive bias, while leaving its interaction with optimization and learned representations open.

The scope of the evaluation suggests several directions for further study. Small-image results use training holdouts, and separating the effects of model width, head count, and rotary dimension would clarify the ImageNet-1K comparisons. Layout gains concern the tested transformations; orientation changes remain a distinct challenge. The masks add no parameters, while the current dense implementation leaves their computational sparsity unexploited. Adaptive head allocation and sparse attention kernels could extend the present study.

6 Conclusion

We introduced RCViT, which represents horizontal and vertical relationships through alternating sparse row–column attention. Deterministic masks compose these relationships across layers, while global heads preserve unrestricted interactions. The construction adds no trainable parameters and is compatible with existing positional encodings.

Classification experiments establish the utility of this spatial structure without positional embeddings and its complementarity with positional encoding under mixed attention. Head-allocation and attention-pattern ablations identify the arrangement and strength of the constraint as important design choices. Resolution and layout evaluations further characterize the resulting generalization across input geometries. Together, these findings support attention connectivity as a means of incorporating spatial relationships into a ViT, with directional and global heads balancing spatial constraints and unrestricted context.

AI USE STATEMENT

Generative AI assisted with language editing, literature and reference checks. This introduced no new experimental measurements. The authors retain responsibility for the final manuscript, including verification of all AI-assisted content.

ETHICS STATEMENT

This work concerns ViT. No new human-subject data were collected. ViT can amplify representation bias, so deployment in consequential settings requires dataset-specific fairness, privacy, and error analysis.

REFERENCES

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9355–9366, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/4e0928de075538c593fbdabb0c5ef2c3-Abstract.html.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3KWnuT-R1bh.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_HJlnC1rKPB.html.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18613–18624, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

Stéphane D'Ascoli, Hugo Touvron, Matthew L. Leavitt, Ari S. Morcos, Giulio Biroli, and Levent Sagun. ConViT: Improving vision transformers with soft convolutional inductive biases. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2286–2296, 2021. URL https://proceedings.mlr.press/v139/d-ascoli21a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. CSWin Transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12124–12134, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Dong_CSWin_Transformer_A_General_Vision_Transformer_Backbone_With_Cross-Shaped_Windows_CVPR_2022_paper.html.

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

Anthony Fuller, Daniel G. Kyrollos, Yousef Yassin, and James R. Green. LookHere: Vision transformers with directed attention generalize and extrapolate. In *Advances in Neural Information Processing Systems*, volume 37, pp. 19683–19739, 2024. doi: 10.52202/079017-0621. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/22d72e9f55bc29cafcca6814a7feac8c-Abstract-Conference.html.

Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer, 2022. URL https://arxiv.org/abs/2209.15001.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6185–6194, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Hassani_Neighborhood_Attention_Transformer_CVPR_2023_paper.html.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, 2024.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers, 2019. URL https://arxiv.org/abs/1912.12180.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 603–612, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Huang_CCNet_Criss-Cross_Attention_for_Semantic_Segmentation_ICCV_2019_paper.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Boyang Li, Yulin Wu, Sizhe Xu, Nuoxian Huang, Zhonghang Yuan, Shangyi Guo, Shu Yang, and Takahiro Yabe. nD-RoPE: A generalized RoPE for n-dimensional position embedding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63283.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL https://research.google/pubs/reading-digits-in-natural-images-with-unsupervised-feature-learning/.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/3416a75f4cea9109507cacd8e2f2aefc-Abstract.html.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1453_ECCV_2020_paper.php.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12894–12904, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Vaswani_Scaling_Local_Self-Attention_for_Parameter_Efficient_Visual_Backbones_CVPR_2021_paper.html.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1564_ECCV_2020_paper.php.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021a.

Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10033–10041, 2021b.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yun_CutMix_Regularization_Strategy_to_Train_Strong_Classifiers_With_Localizable_Features_ICCV_2019_paper.html.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.09412.

A Implementation Details

A.1 Directional head assignment

For the four-head configuration, let $b_p = p \bmod 4$. The directional roles are assigned to physical head indices modulo four:

$$\begin{array}{rcl} b_p & : & \text{orthogonal axis}, \quad \Delta_{\leftrightarrow}, \\ b_p + 1 & : & \text{active axis}, \quad \Delta_{\leftrightarrow}, \\ b_p + 2 & : & \text{active axis}, \quad \Delta_{+}, \\ b_p + 3 & : & \text{active axis}, \quad \Delta_{-}. \end{array} \tag{9}$$

The orthogonal head uses the axis perpendicular to a_ℓ, providing cross-axis interactions within the fully directional block. Mixed attention removes the first $H - k$ masks in the cyclic ordering. For $H = 4$ and $k = 1$, the orthogonal, bidirectional, and forward heads become global, leaving the backward head constrained. Head roles change across layer pairs without moving or tying their learned projection parameters.

A.2 Masks and dilation schedule

Scale schedule. For an axis length $A \geq 2$, define

$$q_A = \max(1, \lfloor A/2 \rfloor), \qquad m_A = \max(1, q_A - \lfloor q_A/3 \rfloor). \tag{10}$$

The six-entry schedule is

$$\boldsymbol{s}(A) = \mathrm{clip}_{[1,A-1]}(1, 2, 4, m_A, A-1, A-1). \tag{11}$$

The layer-pair index selects the same schedule entry on each axis; the numeric dilation can differ on a rectangular grid because the two axis lengths differ. The orthogonal head uses the schedule for its own axis. The twelve-block models use all six layer pairs once.

Mask construction. Masks are reconstructed from valid offsets for each supported patch grid. The query always attends to itself; off-grid keys are discarded. Other connections receive $-\infty$ before softmax, while global heads use zero masks. Reconstructing masks changes neither projection dimensions nor the classifier.

Scope of the permutation argument. Equation 2 follows from $Q(PX) = PQ(X)$, $K(PX) = PK(X)$, and $V(PX) = PV(X)$, together with row-wise softmax commuting with a simultaneous row and column permutation:

$$\begin{aligned} \mathcal{A}(PX; PMP^\top) &= \mathrm{softmax}\Big[P\left(QK^\top/\sqrt{d_h} + M\right)P^\top\Big]\,PV \\ &= P\,\mathrm{softmax}\Big(QK^\top/\sqrt{d_h} + M\Big)\,V. \end{aligned} \tag{12}$$

Without additional positional operations, fixed masks preserve this equivariance when $PM_{\ell h}P^\top = M_{\ell h}$ for every layer and head. This structural condition does not guarantee that learned features distinguish every permutation that alters a mask.

B Training Protocols

All classification models use a feed-forward expansion ratio of four, standard layer normalization, and a single classification objective. There is no auxiliary loss associated with the masks. Small-image experiments use AdamW with peak learning rate 5×10^{-4}, weight decay 0.05, batch size 96, five warmup epochs, and cosine decay over 120 epochs. The held-out training split is identical across allocations. The ImageNet-100 and ImageNet-1K recipes follow the data-efficient ViT training setting (Touvron et al., 2021), without a distillation teacher. They use random resized crops, horizontal flipping, color jitter, RandAugment (Cubuk et al., 2020), random erasing, Mixup (Zhang et al., 2018), CutMix (Yun et al., 2019), and label smoothing. The Mixup and CutMix coefficients are 0.8 and 1.0, label smoothing is 0.1, and EMA decay is 0.9999. The learning rate decays to 10^{-6}, gradient

clipping uses norm 1.0, and stochastic depth is 0.1. ImageNet-100 uses microbatches of 128 with four accumulation steps. ImageNet-1K uses microbatches of 256 with two accumulation steps and repeated augmentation. Thus, both have effective batch size 512.

For ImageNet-1K, ViT-T uses width 192 and either three or four heads, giving head dimensions 64 or 48; ViT-S uses width 384 and six heads, giving head dimension 64. The nD-RoPE rotary dimension is 60 for the three- and six-head models and 48 for the four-head model. Each matched baseline–RCViT pair uses the same rotary dimension. These settings make the four-head comparison a matched architectural configuration, not an isolated change in head count relative to the three-head model.

ImageNet evaluation uses bicubic interpolation and a crop ratio of 0.95, with the same positional interpolation or grid adaptation within each matched pair. ImageNet results use the final-epoch EMA model; small-image results use the best EMA checkpoint on the fixed training holdout.

Table 8: **Classification protocols.** All training inputs are resized or cropped to 224×224 and use 16×16 patches. Holdouts are drawn from the training split, not the standard test sets.

Dataset	Classes	Epochs	Evaluation images	Checkpoint
CIFAR-10	10	120	5,000 (holdout)	Best EMA
CIFAR-100	100	120	5,000 (holdout)	Best EMA
SVHN	10	120	7,328 (holdout)	Best EMA
ImageNet-100	100	400	5,000 (validation)	Final EMA
ImageNet-1K	1,000	300	50,000 (validation)	Final EMA

Small-image training resizes the image before a random crop with four-pixel padding. CIFAR models use horizontal flipping with probability 0.5; SVHN does not use flipping. Evaluation resizes each image without a random crop. These recipes are identical within every baseline–RCViT comparison, including the positional-encoding experiments in Appendix C.2. Table 8 summarizes the evaluation sets and checkpoint rules across the five datasets.

Runtime measurement. Each timed optimizer step accumulates four microbatches of 128 images and includes forward and backward computation, the optimizer update, and the EMA update. Synthetic inputs are generated once on the GPU. Each of five paired trials uses ten warmup steps followed by 50 measured steps, with CUDA synchronization at the timing boundaries. The ratios in Table 7 compare RCViT with its matched positional baseline; they exclude data loading and augmentation.

C Additional Small-Image Experiments

C.1 Head allocation and learning curves

Table 9 gives the full allocation study without positional embeddings. The CIFAR datasets favor four directional heads, while SVHN favors two directional and two global heads. Figure 3 shows that the directional models attain higher accuracy during training as well as at their selected checkpoints.

Table 9: Best-EMA held-out validation top-1 accuracy (%) without positional embeddings. Values are mean $\pm$ sample standard deviation, with five runs for CIFAR-10 and SVHN and three for CIFAR-100. k counts directional heads out of four.

Directional heads k	CIFAR-10	SVHN	CIFAR-100
0	77.50 ± 0.26	67.00 ± 1.95	50.25 ± 0.34
1	88.24 ± 0.37	95.73 ± 0.12	61.21 ± 0.57
2	88.96 ± 0.47	96.11 ± 0.09	62.67 ± 0.41
3	89.41 ± 0.39	96.07 ± 0.22	63.47 ± 0.28
4	89.58 ± 0.11	95.73 ± 0.09	66.43 ± 0.36

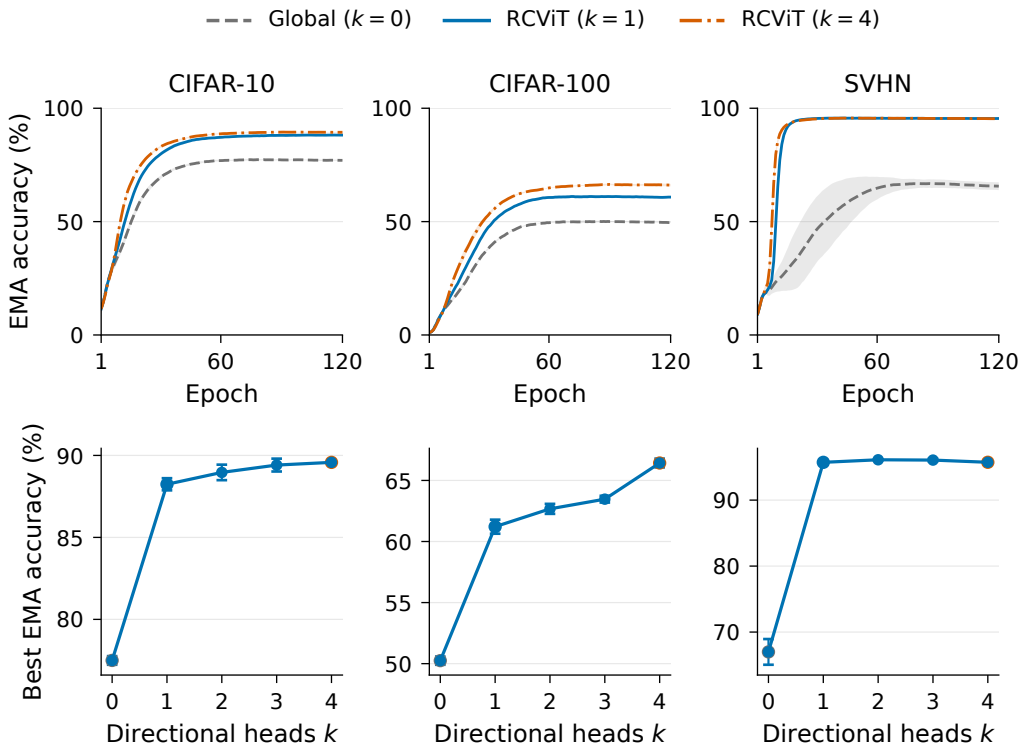

Figure 3: **Learning curves and head allocation without positional embeddings.** Top: EMA validation accuracy at every training epoch for global, mixed, and fully directional attention. Bottom: best-EMA validation accuracy for every allocation of four heads. Lines show means; shaded bands and error bars show sample standard deviations over five runs for CIFAR-10 and SVHN and three for CIFAR-100. No smoothing is applied. Learning curves show the current EMA model at each epoch, whereas the allocation results use each run's best EMA checkpoint.

The learning trajectories distinguish improvements at the selected checkpoint from differences sustained across training. On SVHN, mixed and fully directional attention separate from the global baseline early and remain close to each other thereafter. On the CIFAR datasets, the fully directional model eventually attains a higher mean accuracy than the mixed model. SVHN instead favors two directional heads. Together, the learning and allocation curves indicate that the preferred balance between directional and global attention depends on the dataset.

C.2 Combining directional attention with positional encoding

Table 10 extends the small-image study to four positional mechanisms. Each comparison retains the same positional operation and replaces global attention with four directional heads. All twelve mean differences are positive, but their magnitudes vary substantially. The CIFAR-100 comparisons show the largest gains, while SVHN with nD-RoPE is essentially unchanged, with a paired difference of 0.01 ± 0.07 percentage points. The contribution of directional attention thus varies with the existing positional mechanism and dataset.

These findings complement the ImageNet-100 study. On the small-image tasks, positional encoding does not eliminate the benefit of restricting every head. On ImageNet-100, retaining global heads is more effective: global baselines exceed the fully directional models, and mixed attention improves on both. The datasets and training recipes differ, so the comparison motivates task-dependent constraint strength without attributing the difference to dataset size alone.

Table 10: **Positional encoding with fully directional attention.** Best-EMA holdout top-1 accuracy (%), mean ± sample standard deviation over five runs for CIFAR-10 and SVHN and three for CIFAR-100. The final column summarizes paired accuracy differences in percentage points.

Dataset	Position	Global ($k = 0$)	RCViT ($k = 4$)	Paired Δ (pp)
CIFAR-10	Learned APE	84.88±0.30	89.49±0.23	+4.61±0.42
	Relative bias	81.72±0.58	89.34±0.24	+7.62±0.73
	CPVT-CPE	84.79±0.27	89.49±0.18	+4.70±0.24
	nD-RoPE	85.47±0.29	89.15±0.24	+3.68±0.32
CIFAR-100	Learned APE	56.89±0.34	66.89±0.24	+10.00±0.47
	Relative bias	52.89±1.29	65.86±0.19	+12.97±1.47
	CPVT-CPE	57.87±1.08	66.59±0.36	+8.71±0.76
	nD-RoPE	58.95±0.16	65.67±0.37	+6.72±0.30
SVHN	Learned APE	92.27±0.90	95.81±0.09	+3.54±0.84
	Relative bias	94.39±0.20	95.60±0.19	+1.22±0.15
	CPVT-CPE	90.99±0.54	95.75±0.11	+4.77±0.51
	nD-RoPE	95.46±0.15	95.48±0.09	+0.01±0.07

C.3 Qualitative comparisons across datasets

Figure 4 compares the no-position classifiers on three holdouts, including a reverse case for each dataset. These examples illustrate differences in the learned classifiers, not attention-based explanations. Table 9 reports the aggregate results.



Figure 4: **Classification examples on three small-image datasets.** Actual evaluation inputs for one matched best-EMA pair per dataset, without positional embeddings. Blue denotes correct predictions; orange denotes errors. In fixed holdout order, we select the first two RCViT-only correct cases and the first global-only correct case with distinct true classes. The selection illustrates disagreement, not its frequency.

D Additional ImageNet Analyses

D.1 Training dynamics on ImageNet-100

Figure 5 shows the complete 400-epoch learning curves for global attention, one directional head, and four directional heads under each positional mechanism. The mixed models attain the highest mean accuracy at the end of training, followed by global and fully directional attention. The later-epoch views show the benefit of retaining global interactions without selecting an intermediate checkpoint. With nD-RoPE, the global and mixed models have nearly equal mean accuracy at the midpoint of training, before separating later. Thus, an early checkpoint would not fully reflect the final comparison in this setting.

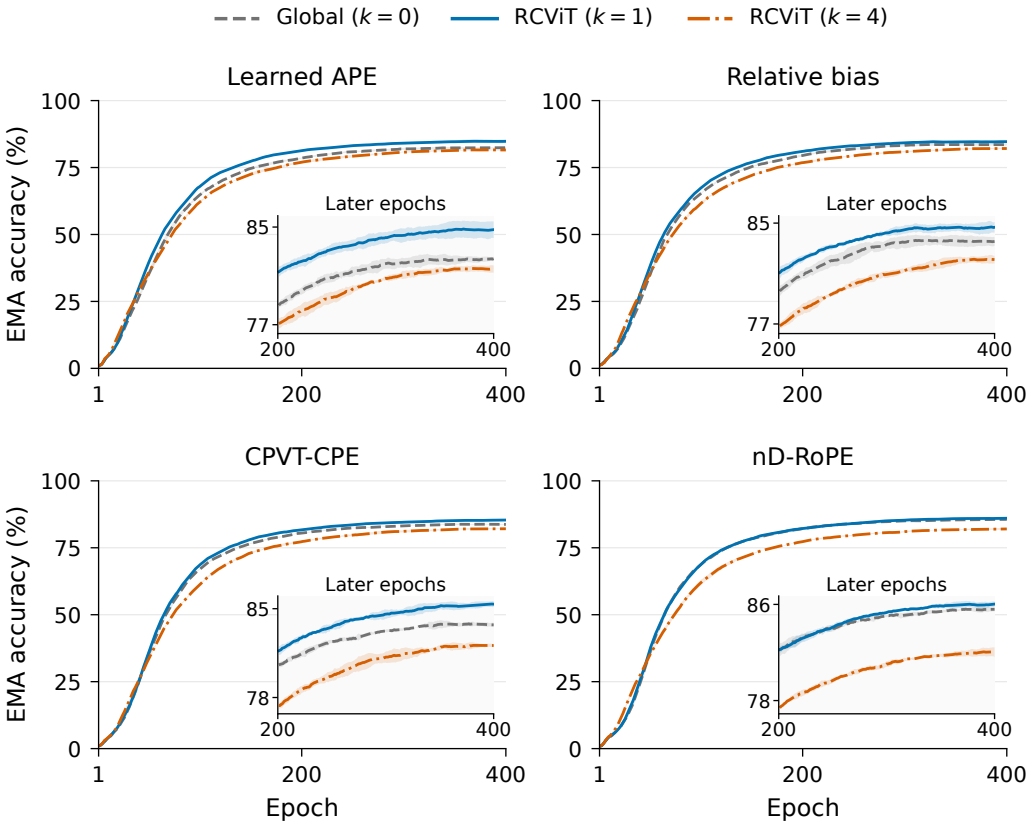

Figure 5: **ImageNet-100 learning curves with four positional mechanisms.** Mean per-epoch EMA validation top-1 accuracy over three matched runs; shaded bands show sample standard deviation. Insets show the second half of training on expanded vertical scales. All curves use the same 400-epoch duration without smoothing or best-checkpoint selection. Global and fully directional attention are compared with the mixed $k = 1$ configuration used in the main results.

D.2 Training dynamics on ImageNet-1K

Figure 6 reports the 300-epoch curves for each backbone and positional configuration in Table 3. The learned-APE ViT-T and nD-RoPE ViT-S comparisons favor RCViT both at the midpoint and at the end of training. The two nD-RoPE ViT-T comparisons favor global attention at these points. The learning curves therefore show the same configuration-dependent pattern as the final evaluation. These are matched architectural comparisons: the four-head model also has a different head dimension and rotary dimension, as specified in Appendix B.

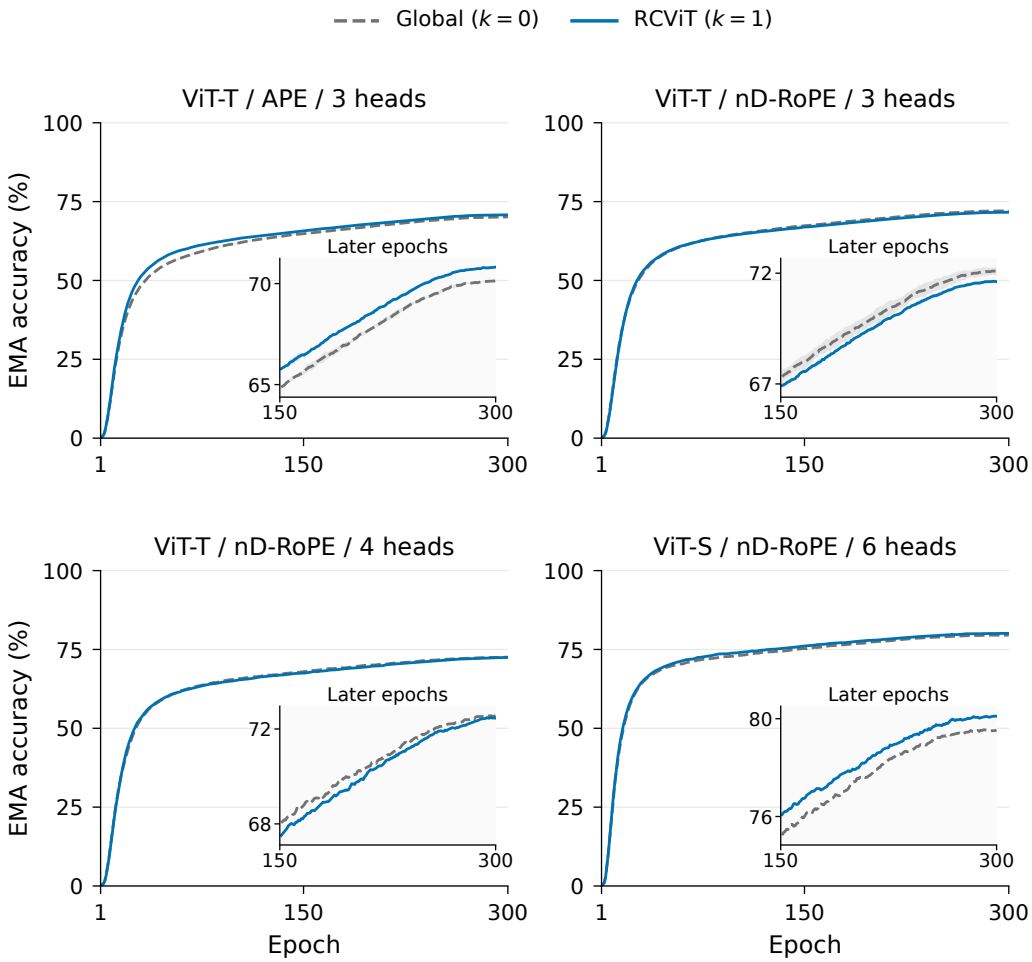

Figure 6: **ImageNet-1K learning curves across model configurations.** Per-epoch EMA validation top-1 accuracy for the matched comparisons in Table 3. The top row reports means and sample standard deviations over three runs; the bottom row reports one matched pair per configuration and has no uncertainty band. Insets show the second half of training. Curves use the same run set throughout training and are not smoothed. RCViT constrains one head in each configuration.

D.3 Sensitivity to global-head allocation

Figure 7 expresses the complete ImageNet-100 allocation study as paired changes from the global baseline. Increasing the number of global heads from zero to one improves every positional configuration, but the strongest allocation depends on the positional mechanism. Three global heads improve over global attention for all four mechanisms. Two global heads improve the APE, relative-bias, and CPVT-CPE models, but are approximately equal to the nD-RoPE baseline. The fully global endpoint has zero difference by definition and serves as the within-run reference, not as an additional independent measurement. These paired curves complement the absolute accuracies in Table 4.

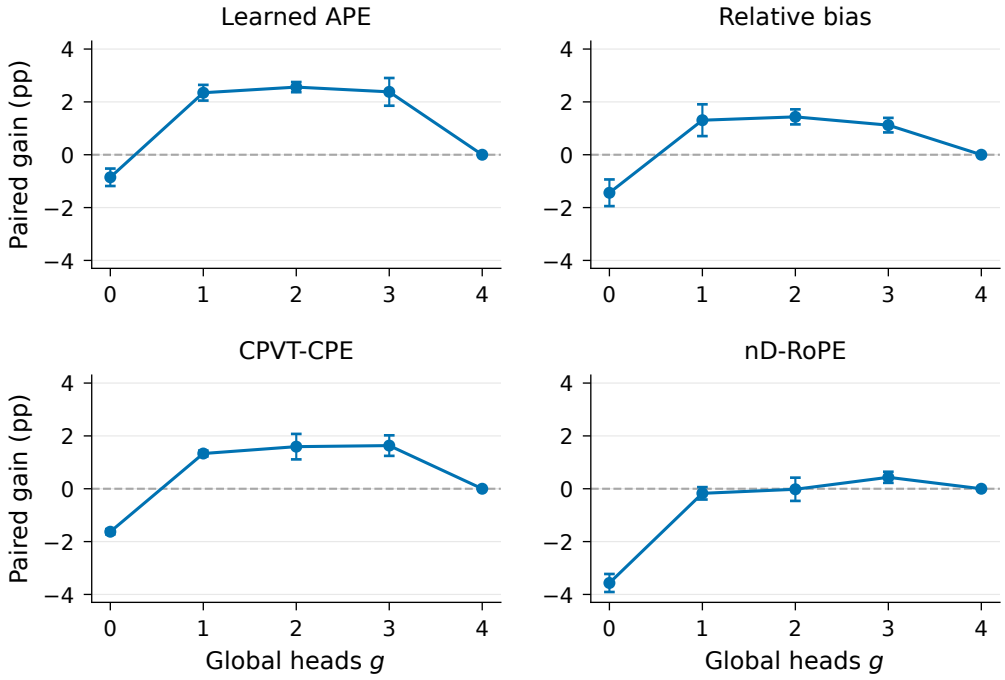

Figure 7: **Accuracy changes across global-head allocations.** ImageNet-100 final-EMA accuracy relative to the matched global baseline, in percentage points. Each model has four heads, with $k = 4 - g$ directional heads. Points and error bars show the mean and sample standard deviation of paired differences over three runs. The dashed line indicates no change from global attention.

D.4 Construction of attention-pattern controls

The controls in Table 5 modify only the single constrained head; the other three heads remain global. Random and local neighborhoods match the number of admissible keys for each query in the directional reference and preserve the self-connection. Random neighborhoods select the remaining keys without replacement and remain fixed during training. Local neighborhoods select the nearest patches by squared Euclidean distance on the grid, breaking ties by raster-order index. These controls distinguish the arrangement of connections from simply reducing their number.

The row-only and column-only controls retain the offset construction but use one axis throughout the network. The fixed-head control retains the alternating masks and applies them to the same physical head in every layer, while the standard configuration changes the head assignment across layer pairs. The latter comparison changes head assignment, not the row–column neighborhoods. Both assignments improve on the global reference, with a slightly higher mean for the fixed assignment; the shared benefit is associated with axis alternation rather than a particular head-assignment schedule.

E Spatial Transformation Evaluation

The square resolutions are 160, 192, 224, 256, 288, 320, 384, 448, and 512 pixels. Rectangular inputs use 224×336, 336×224, 256×384, and 384×256. Canvas tests place an unchanged 224×224 evaluation image at the left, center, or right of a 224×336 canvas; at the top, center, or bottom of a 336×224 canvas; or at the top-left, center, or bottom-right of a 320×448 canvas. A clean 224×224 canvas provides a consistency reference. Additional tests translate the image by ± 16 pixels along each axis, transpose its spatial dimensions, or rotate it clockwise by 90 degrees. Together, these define 29 evaluations, including the two clean references, for every checkpoint.

For transformation group T and training run i, we first compute the mean accuracy $a_{i,T} = |T|^{-1} \sum_{t \in T} a_{i,t}$. Table 6 reports the mean and sample standard deviation of $a_{i,T}$ across runs.

Clean-relative drop is $d_{i,T} = a_{i,\text{clean}} - a_{i,T}$. Table 11 reports the paired difference in this drop between RCViT and its global baseline. Negative values indicate a smaller drop with RCViT; positive values indicate a larger drop. This metric distinguishes the transformation response from an improvement already present on clean images.

Table 11: Change in accuracy drop relative to clean inputs, RCViT minus global baseline (percentage points). Mean ± sample standard deviation of paired differences over three runs. Negative values denote a smaller drop. The transformation groups match Table 6.

Position	Resolution	Aspect ratio	Canvas
Learned APE	-0.95 ± 0.21	-0.30 ± 0.24	-1.16 ± 0.44
Relative bias	-0.58 ± 0.37	-0.45 ± 0.51	-0.34 ± 0.27
CPVT-CPE	0.19 ± 0.31	0.16 ± 0.31	0.18 ± 0.39
nD-RoPE	-0.17 ± 0.36	0.34 ± 0.45	-0.09 ± 0.74

Table 12: Additional translation and orientation tests on ImageNet-100. Top-1 accuracy (%), mean ± sample standard deviation over three runs. Translation averages four ±16-pixel horizontal or vertical shifts within each run. Transpose and clockwise 90-degree rotation are single transformations.

Position	Attention	Translation	Transpose	Rotation
Learned APE	Global	82.22 ± 0.29	58.45 ± 0.30	58.87 ± 0.20
	+ RCViT	84.53 ± 0.58	58.23 ± 0.53	57.99 ± 0.43
Relative bias	Global	83.52 ± 0.25	59.02 ± 0.95	58.74 ± 0.90
	+ RCViT	84.57 ± 0.23	59.05 ± 0.45	59.12 ± 0.44
CPVT-CPE	Global	83.58 ± 0.06	58.82 ± 0.87	58.97 ± 0.88
	+ RCViT	85.08 ± 0.18	58.61 ± 0.25	58.41 ± 0.27
nD-RoPE	Global	85.34 ± 0.06	58.68 ± 0.06	58.56 ± 0.26
	+ RCViT	85.83 ± 0.13	58.54 ± 0.42	58.61 ± 0.47

Table 12 shows improvements on translated inputs across the four positional mechanisms, while transposition and rotation yield mixed changes. These transformations alter the orientation of image content relative to the directional neighborhoods. The architecture specifies row–column relationships without imposing orientation invariance.

E.1 Results for every evaluation geometry

Figure 8 reports the mean paired accuracy change for each non-clean transformation, rather than averaging over a transformation group. Tables 13 and 14 provide both models' absolute accuracies and run-to-run variation for all 29 evaluations. Each entry uses the same 5,000 ImageNet-100 validation images. The two clean references use the same input geometry and provide consistent results.

This view distinguishes resolution, placement, and orientation effects. Accuracy gains under square resizing extend across the tested resolution range, while the sign under transposition and rotation depends on the positional mechanism. Positive cells measure higher transformed accuracy, not necessarily a smaller loss from clean inputs; the clean-relative comparisons in Table 11 address that separate question. None of these evaluations changes the trained projection parameters or fine-tunes the classifier for a new geometry.

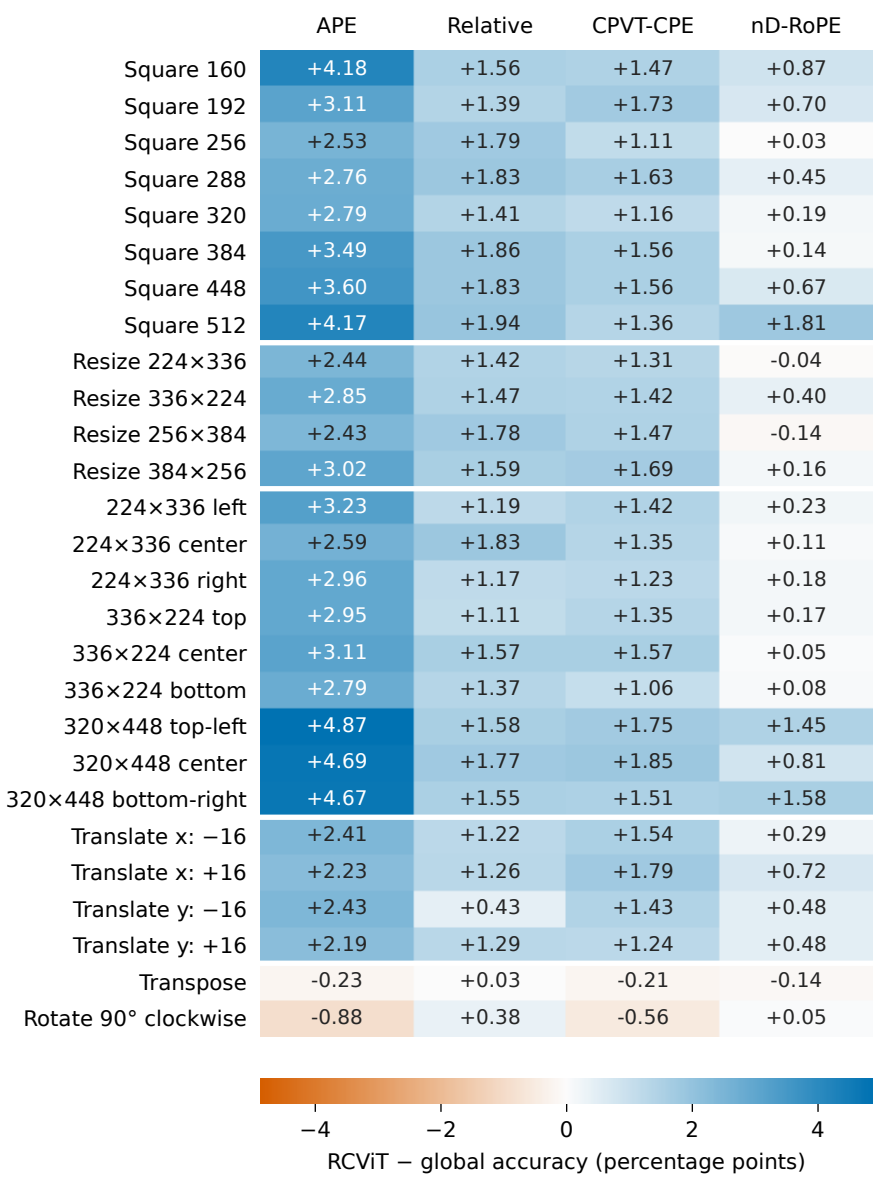

Figure 8: **Paired accuracy changes for individual spatial transformations.** Each cell is the mean RCViT-minus-global difference over three matched ImageNet-100 runs, in percentage points. Blue denotes higher accuracy with RCViT and orange denotes lower accuracy. The two clean references are omitted from this plot but retained in Tables 13 and 14. Canvas labels give dimensions and image placement; resizing changes image dimensions, whereas canvas padding leaves the crop unchanged.

Table 13: **Complete square and rectangular resolution evaluation.** ImageNet-100 top-1 accuracy (%), mean ± sample standard deviation over three runs. Each input has paired rows for global attention ($k = 0$) and RCViT ($k = 1$). Square inputs are specified by side length; rectangular dimensions are height × width in pixels.

Input	k	Learned APE	Relative bias	CPVT-CPE	nD-RoPE
Square resizing					
160	0	72.34±0.51	75.56±0.61	76.05±0.23	78.31±0.15
	1	76.52±0.53	77.12±0.35	77.51±0.28	79.17±0.34
192	0	79.21±0.34	80.85±0.50	81.01±0.07	82.81±0.01
	1	82.33±0.57	82.23±0.57	82.73±0.19	83.51±0.22
224	0	82.41±0.16	83.55±0.35	83.73±0.17	85.59±0.17
	1	84.79±0.68	84.67±0.42	85.36±0.23	86.03±0.24
256	0	83.51±0.73	83.97±0.61	84.83±0.21	86.92±0.16
	1	86.04±0.45	85.76±0.14	85.94±0.28	86.95±0.18
288	0	83.53±0.32	84.25±0.18	85.10±0.22	87.39±0.37
	1	86.29±0.35	86.08±0.21	86.73±0.06	87.84±0.24
320	0	83.75±0.53	84.62±0.05	85.46±0.09	87.55±0.17
	1	86.55±0.43	86.03±0.11	86.62±0.14	87.74±0.07
384	0	82.59±0.67	84.17±0.31	84.89±0.14	87.16±0.12
	1	86.08±0.40	86.03±0.24	86.45±0.44	87.30±0.15
448	0	81.34±0.95	82.96±0.55	83.83±0.16	85.18±0.42
	1	84.94±0.29	84.79±0.16	85.39±0.28	85.85±0.27
512	0	79.46±0.53	81.57±0.55	82.43±0.17	82.41±0.69
	1	83.63±0.45	83.51±0.16	83.79±0.22	84.22±0.16
Rectangular resizing					
224×336	0	83.32±0.55	83.81±0.47	84.53±0.12	86.87±0.16
	1	85.76±0.70	85.23±0.16	85.83±0.33	86.83±0.13
336×224	0	82.11±0.14	83.14±0.45	83.13±0.20	85.61±0.14
	1	84.96±0.33	84.61±0.20	84.55±0.24	86.01±0.15
256×384	0	83.21±0.55	84.04±0.40	84.73±0.26	87.20±0.23
	1	85.64±0.41	85.82±0.11	86.19±0.19	87.06±0.12
384×256	0	82.04±0.07	83.39±0.34	83.43±0.17	86.01±0.22
	1	85.06±0.47	84.98±0.24	85.11±0.26	86.17±0.29

With nD-RoPE, the wider rectangular inputs remain close to the global baseline, with mean differences of -0.04 and -0.14 percentage points, while the taller inputs favor RCViT. These shape-dependent differences complement the positive group average. Figure 9 further separates these effects from the gain already present on clean inputs.

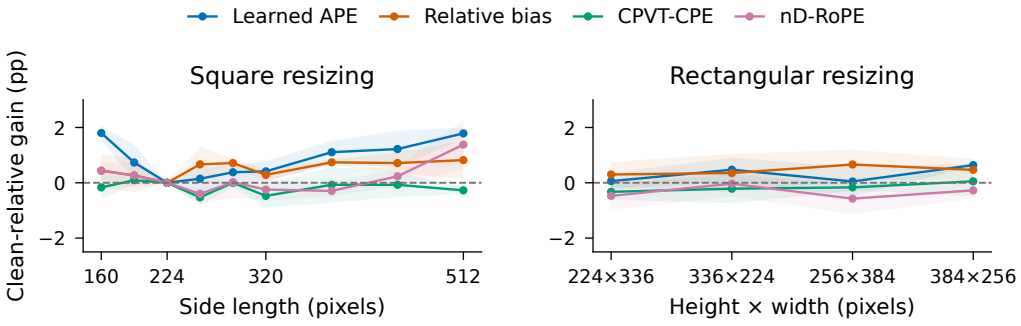

Figure 9: **Resolution effects after accounting for clean accuracy.** Paired gains after subtracting the gain at 224 × 224. Positive values indicate a smaller clean-relative accuracy drop with RCViT. Lines and bands show means and sample standard deviations over three runs.

Table 14: **Complete canvas, translation, and orientation evaluation.** ImageNet-100 top-1 accuracy (%), mean ± sample standard deviation over three runs. Paired rows compare global attention ($k = 0$) and RCViT ($k = 1$). Canvas inputs retain an unchanged 224×224 crop; TL and BR denote top-left and bottom-right placements. CW denotes clockwise rotation.

Input	k	Learned APE	Relative bias	CPVT-CPE	nD-RoPE
Padded canvas					
224×224 clean	0	82.41±0.16	83.55±0.35	83.73±0.17	85.59±0.17
	1	84.79±0.68	84.67±0.42	85.36±0.23	86.03±0.24
224×336 left	0	80.43±0.42	81.65±0.39	82.79±0.20	84.96±0.42
	1	83.66±0.63	82.85±0.17	84.21±0.36	85.19±0.32
224×336 center	0	81.01±0.61	81.02±0.29	82.39±0.11	84.57±0.27
	1	83.61±0.31	82.85±0.01	83.74±0.16	84.69±0.31
224×336 right	0	80.56±0.37	81.71±0.49	82.83±0.22	84.90±0.26
	1	83.52±0.47	82.87±0.30	84.06±0.27	85.08±0.36
336×224 top	0	80.32±0.24	81.59±0.67	82.71±0.22	84.83±0.36
	1	83.27±0.75	82.70±0.39	84.05±0.46	85.00±0.33
336×224 center	0	80.08±0.45	81.05±0.71	81.93±0.25	84.63±0.18
	1	83.19±0.41	82.62±0.37	83.50±0.40	84.67±0.08
336×224 bottom	0	80.53±0.49	81.57±0.59	82.63±0.30	84.75±0.21
	1	83.33±0.68	82.94±0.45	83.69±0.25	84.83±0.17
320×448 TL	0	73.87±0.44	75.43±1.25	78.13±0.64	80.06±1.27
	1	78.75±0.14	77.01±0.29	79.88±1.01	81.51±0.53
320×448 center	0	74.42±0.33	75.26±1.11	77.59±0.40	80.99±0.58
	1	79.11±0.29	77.03±0.56	79.43±0.81	81.80±0.40
320×448 BR	0	74.25±0.36	75.78±1.11	78.39±0.60	79.72±1.09
	1	78.91±0.25	77.33±0.87	79.89±0.65	81.30±0.30
Translation					
x: −16	0	82.11±0.27	83.35±0.51	83.59±0.05	85.35±0.04
	1	84.52±0.59	84.57±0.24	85.13±0.25	85.64±0.13
x: +16	0	82.47±0.39	83.55±0.06	83.43±0.09	85.40±0.07
	1	84.71±0.73	84.81±0.20	85.22±0.07	86.12±0.12
y: −16	0	82.14±0.24	83.80±0.28	83.49±0.20	85.28±0.07
	1	84.57±0.53	84.23±0.10	84.91±0.19	85.76±0.14
y: +16	0	82.15±0.36	83.39±0.25	83.81±0.18	85.34±0.27
	1	84.35±0.51	84.69±0.43	85.05±0.22	85.82±0.16
Orientation					
Transpose	0	58.45±0.30	59.02±0.95	58.82±0.87	58.68±0.06
	1	58.23±0.53	59.05±0.45	58.61±0.25	58.54±0.42
Rotate 90° CW	0	58.87±0.20	58.74±0.90	58.97±0.88	58.56±0.26
	1	57.99±0.43	59.12±0.44	58.41±0.27	58.61±0.47

In the canvas evaluations, all four positional mechanisms improve at every tested placement. The paired examples below illustrate individual prediction differences in both directions alongside these aggregate gains.

E.2 Qualitative examples under a change in canvas geometry

Figure 10 changes canvas geometry using mean-filled padding without altering image content. After padding, only RCViT correctly classifies the chime and borzoi, whereas only the global baseline correctly classifies the English foxhound. These examples illustrate variation in image-level responses alongside the aggregate gains in Table 6.

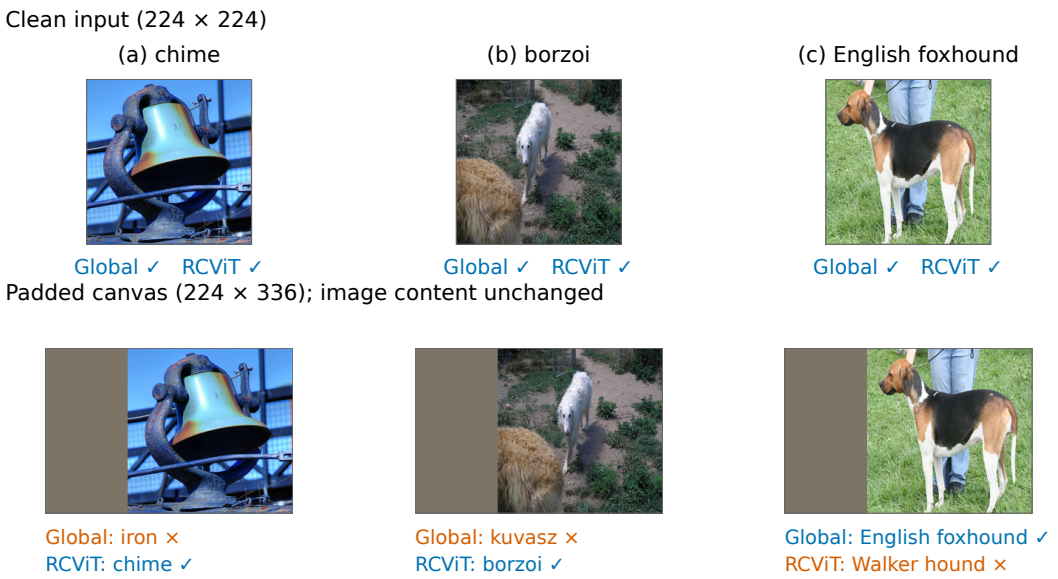

Figure 10: **Predictions under a change in canvas geometry.** Original crops (top) and right-aligned padded inputs (bottom) are evaluated with one matched pair of final-EMA learned-APE models. Headings give ground-truth classes; check marks indicate correct predictions. In validation order, we select the first two cases favoring RCViT and the first reverse case from distinct classes, conditional on both models being correct before padding. Table 6 reports the full-set results.

F Additional Computational Measurements

Table 15 provides the absolute optimizer-step times underlying Table 7, together with peak allocated GPU memory. Timing uncertainty is computed across five trial means, each of which averages 50 synchronized steps after warmup. Memory is the largest peak allocated value across the five trials, rather than reserved GPU memory. These measurements use the same device, microbatch size, gradient accumulation, and optimizer operations within each pair.

Table 15: **Absolute optimizer-step time and peak allocated memory.** Measurements on an NVIDIA RTX 5880 Ada Generation GPU with four accumulated microbatches of 128 GPU-resident images. Times are mean ± sample standard deviation across five trials; memory is the maximum peak allocation across those trials. Forward and backward computation, optimizer updates, and EMA updates are included. Data loading and augmentation are excluded.

Position	Attention	Optimizer step (ms)	Peak memory (GiB)
Learned APE	Global	243.62±9.52	2.528
	+ RCViT	220.38±39.38	2.529
Relative bias	Global	238.36±16.20	3.855
	+ RCViT	248.12±15.82	3.847
CPVT-CPE	Global	245.83±17.07	2.547
	+ RCViT	224.77±44.85	2.529
nD-RoPE	Global	271.94±12.72	3.855
	+ RCViT	283.88±17.45	3.847

Peak allocated memory is similar within each pair. The timing estimates describe optimizer-step cost together with its measured variability. Table 7 averages paired ratios, which need not equal the ratio of mean times reported here. These measurements characterize optimizer-step cost, not end-to-end training acceleration.