VISION TRANSFORMER FLOYD: RELATION-CENTRIC VISUAL BACKBONES WITH PIVOTAL ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As vision models are increasingly applied to complex real-world tasks, visual understanding must extend beyond recognizing individual objects to modeling their relationships, interactions, and scene dynamics. Mainstream visual backbones typically operate on patches, leaving relations between patches implicit in the resulting token representations. Pivotal Attention (PA) makes relations explicit through ordered-pair features and their updates, but its all-pairs computation is costly for images. We introduce Row Column Pivotal Attention (RCPA), which exploits the organized image grid to jointly select and aggregate row and column features. This structure preserves global two-hop coverage through sparse connections, as characterized by the rook's graph, while reducing attention interactions from PA's $O(T^3)$ to $O(T^{3/2})$ for T patches on a square grid. Building on RCPA, we develop Vision Transformer Floyd (ViF), a general-purpose backbone that brings PA-inspired relation modeling to visual recognition and dense prediction. On ClockBench relative-angle prediction, PA and RCPA Small reduce IID angular error by 83.6% and 67.3% relative to ViT Small. On real images, ViF Base improves VG150 SGDet R@100 by 4.6% relative to ViT Base and transfers effectively to ImageNet-1K, ADE20K, and COCO. These results establish a practical route from explicit relation modeling to visual backbone design and motivate its extension to richer relational and multimodal tasks.

1 INTRODUCTION

Advances in visual recognition have enabled vision models to support a growing range of tasks in complex real-world environments. These applications demand representations that go beyond object identity to capture relationships, interactions, and scene dynamics. For example, scene graph generation captures subject–predicate–object structures (Im et al., 2024; Kim et al., 2024a; Li et al., 2024b); embodied intelligence uses spatial relationships to ground instructions and actions (Gu et al., 2024; Zhen et al., 2024; Qu et al., 2025); and world models capture interactions to predict scene dynamics (Sehgal et al., 2024; Min et al., 2024; Zhou et al., 2025). In Figure 1(a), identifying the grapes whose color is queried requires first locating the bowl containing the banana: relations determine which object matters.

Mainstream visual backbones model relations *implicitly*: convolution and self-attention exchange information while retaining one feature per location or token (Wang et al., 2018; Dosovitskiy et al., 2021)(Figure 1(b)). Pairwise attention scores guide aggregation, intertwining object attributes and relational information in the outputs. Recent work begins to address this distinction: Abstractors and Dual Attention Transformers (Altabaa et al., 2024; Altabaa & Lafferty, 2025) separate relational and sensory processing. However, their outputs remain object-indexed; directly representing and refining relations themselves presents a further opportunity for backbone design.

Pivotal Attention (PA), introduced by FloydNet (Yu et al., 2026) for graph prediction and algorithmic reasoning, provides a concrete approach to *explicit* relation modeling. PA represents each ordered relation by a learned feature $\mathcal{R}_{ik} \in \mathbb{R}^C$ (Figure 1(c)). To refine it, PA forms candidates from relations $\mathcal{R}_{ij}$ and $\mathcal{R}_{jk}$ through a shared intermediate j. Both source features determine each candidate's score and contribution, and attention over j integrates them into the target relation. Relations are thus refined through other relations, providing a direct foundation for explicit visual relation modeling.

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.

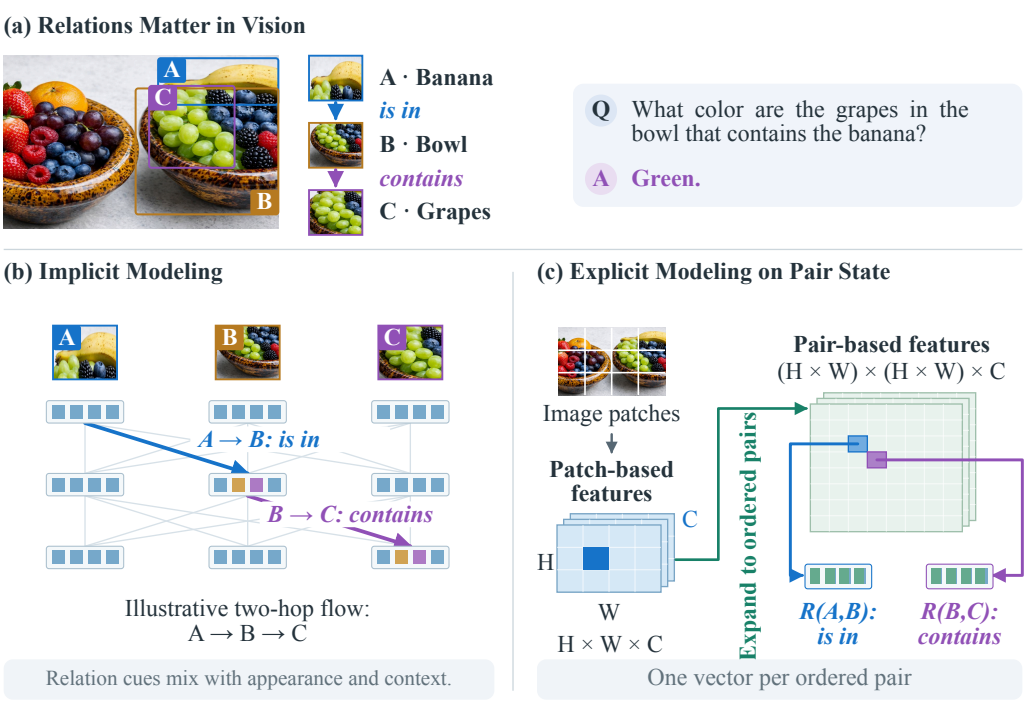


Figure 1: **From visual relationships to explicit relation modeling.** (a) Relations link the banana, bowl, and grapes to identify the queried object. (b) Self-attention updates token features through pairwise scores. (c) PA maintains ordered-pair features and refines them through a shared intermediate. For $T = HW$ patches of dimension C, token and pair states contain TC and T^2C elements.

The obstacle is computational: for T image patches with feature dimension C, PA maintains $O(T^2C)$ pair features and requires $O(T^3)$ attention interactions. Images offer structure that general graphs do not prescribe: patches occupy an ordered grid with natural row and column coordinates. These axes connect any two patches through at most one intermediate location, suggesting a spatial counterpart to PA's intermediate-based updates. We explore this correspondence by pairing row and column features through a shared coordinate index and using them jointly to update a target patch.

We introduce Row Column Pivotal Attention (RCPA) and build Vision Transformer Floyd (ViF)[1] around it. RCPA jointly scores each feature pair and aggregates candidates with a single attention distribution. On square grids, it reduces PA's feature storage to $O(TC)$ and attention interactions to $O(T^{3/2})$. The rook's graph explains the structural choice: uniform row and column connections achieve global two-hop coverage with an asymptotically optimal order of maximum degree. Joint normalization complements this connectivity by coupling both branches within each update, making PA-inspired relation modeling feasible at image-token scale.

To evaluate relation modeling in a controlled setting, we construct ClockBench, a synthetic dataset for predicting directed relative angles between clock hands. PA and RCPA Small reduce IID angular error by 83.6% and 67.3%, respectively, relative to ViT Small. We then evaluate natural-image relation modeling on VG150, a standard subset of Visual Genome (Krishna et al., 2017; Xu et al., 2017). ViF improves scene graph detection across Tiny, Small, and Base scales, with a 4.6% relative gain in Base-scale R@100 over ViT. Finally, ImageNet-1K classification (Russakovsky et al., 2015), ADE20K segmentation (Zhou et al., 2017), and COCO detection and instance segmentation (Lin et al., 2014) demonstrate the utility of the RCPA-based backbone across general visual tasks.

- **PA-inspired visual relation modeling.** We explore explicit relation modeling as a design principle for general-purpose visual backbones, emphasizing relations alongside visual entities as fundamental targets of representation learning.

[1]Floyd reflects the origin in PA, whose intermediate updates are inspired by the Floyd–Warshall algorithm.

- **A computationally feasible, principled design.** We propose RCPA and Vision Transformer Floyd (ViF). RCPA reduces PA's attention interaction complexity from $O(T^3)$ to $O(T^{3/2})$ for T patches on a square grid.
- **Synthetic and real-image validation.** PA and RCPA substantially reduce prediction error compared with conventional backbones on a synthetic dataset for learning geometric relations. Further, ViF outperforms our ViT baselines in scene graph detection and achieves competitive performance on classification, segmentation, and detection tasks.

2 PRELIMINARIES

2.1 PIVOTAL ATTENTION

For T elements with feature dimension C, self-attention maintains token features $X \in \mathbb{R}^{T\times C}$, whereas PA maintains an ordered-pair tensor $\mathcal{R} \in \mathbb{R}^{T\times T\times C}$ (Yu et al., 2026). To update (i,k), it combines (i,j) and (j,k) through a shared intermediate, or *pivot*, j. For head a of dimension d, the target produces a query and the two source pairs produce a candidate key and value:

$$q^a_{ik} = W^a_Q \mathcal{R}_{ik}, \tag{1}$$

$$\kappa^a_{ijk} = W^a_{K,L}\mathcal{R}_{ij} + W^a_{K,R}\mathcal{R}_{jk}, \qquad \nu^a_{ijk} = W^a_{V,L}\mathcal{R}_{ij} + W^a_{V,R}\mathcal{R}_{jk}. \tag{2}$$

Attention selects and aggregates these candidates over pivots:

$$\alpha^a_{ijk} = \mathrm{softmax}_j\left(\frac{(q^a_{ik})^\top \kappa^a_{ijk}}{\sqrt{d}}\right), \qquad o^a_{ik} = \sum_{j=1}^{T} \alpha^a_{ijk}\nu^a_{ijk}. \tag{3}$$

Both source relations influence the attention weight and the aggregated value. PA requires $O(T^2C)$ maintained features and $O(T^3)$ interactions; fused implementations can avoid cubic intermediate storage but retain cubic computation.

2.2 ROOK'S GRAPH

A rook's graph $G_{H,W}$ has one vertex per location on an $H \times W$ grid; two distinct vertices are adjacent exactly when they share a row or column (Brouwer & Haemers, 2012). Figure 2(a) illustrates this support. Locations $u = (i,k)$ and $v = (p,q)$ in different rows and columns communicate through either cross-corner vertex, (i,q) or (p,k); same-row or same-column pairs are directly adjacent.

An equivalent view is $G_{H,W} \cong L(K_{H,W})$, the line graph of a complete bipartite graph (Ducey et al., 2015). Its two parts represent image rows and columns, and each edge (i,k) represents the patch at their intersection. Edges sharing a row or column endpoint correspond to adjacent rook vertices. For u and v above, the two rows and columns induce the four-edge $K_{2,2}$ in Figure 2(b); its colored edges identify the cross-corner patches.

This suggests a spatial alternative to Full PA's explicit state for every ordered patch pair. With $T = HW$, the bipartite view indexes only T patch features, while row–column composition retains two-hop access between arbitrary locations. It motivates adapting PA's paired-source update to this support with $O(TC)$ maintained features, rather than $O(T^2C)$ pair states. We use this structural prior to guide the design of RCPA, adapting PA's paired-source computation to image patches through joint scoring and aggregation of row and column features.

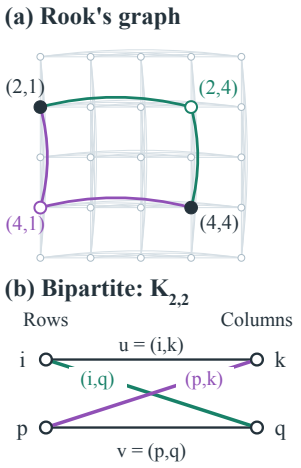


Figure 2: **Rook and bipartite views of the image grid.** Row–column connectivity links any two image patches within two hops.

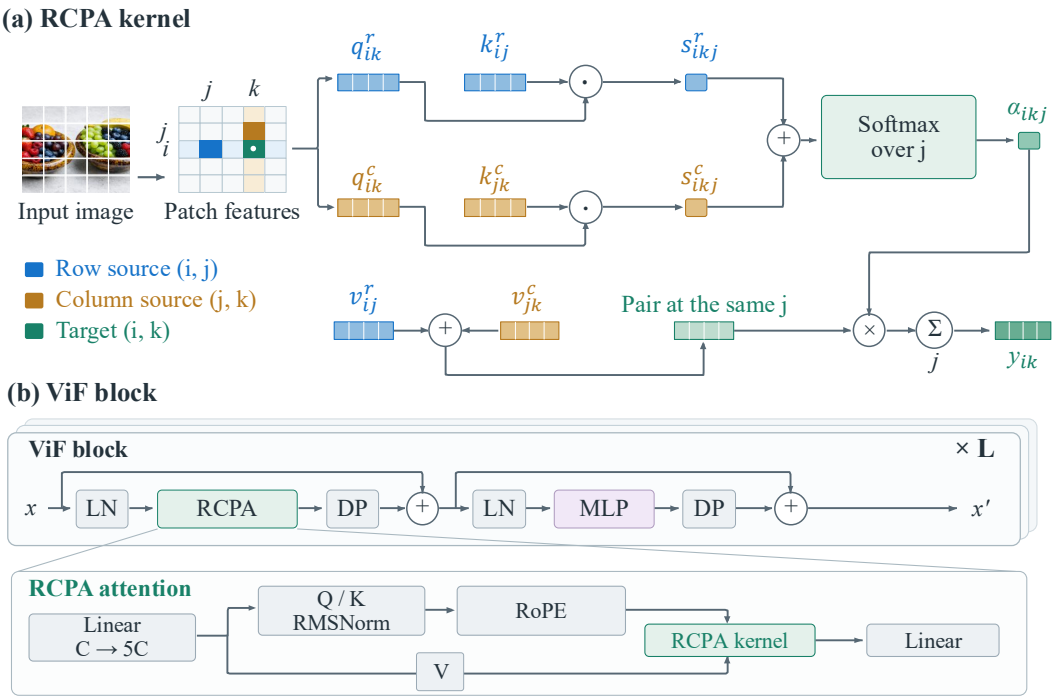


Figure 3: **RCPA and its integration into ViF.** (a) Row and column sources sharing index j receive a joint attention score and a shared aggregation weight. (b) The RCPA module replaces self-attention within a pre-norm Transformer block. Tensor entries are schematic.

3 METHOD

3.1 ROW COLUMN PIVOTAL ATTENTION

RCPA models interactions between row and column features through a joint attention update. For a target patch at (i, k) on an $N \times N$ grid, each index j defines a candidate consisting of the row feature x_{ij} and the column feature x_{jk}. The two sources receive one joint score and one shared weight: both row and column features determine their contribution to the target patch. This adapts PA's paired-source computation to patch features, with j indexing aligned coordinates on the two spatial axes. Figure 3 shows the operator and its integration into a Transformer block.

Given $X \in \mathbb{R}^{H\times W\times C}$, a linear projection produces the query and branch-specific keys and values:

$$Q, K^r, K^c, V^r, V^c \in \mathbb{R}^{M\times H\times W\times d}, \tag{4}$$

for M heads of dimension $d = C/M$; the batch axis is omitted. Superscripts r and c denote row and column branches. After query–key normalization and axis-specific rotary position embedding (RoPE) (Su et al., 2021), the shared learned query projection yields positional versions q^r and q^c for the two branches; the keys below likewise denote their normalized, rotated versions. On a square grid $H = W = N$, target location (i, k) receives two scores for each index j:

$$s^r_{ikj} = \langle q^r_{ik}, k^r_{ij}\rangle, \qquad s^c_{ikj} = \langle q^c_{ik}, k^c_{jk}\rangle. \tag{5}$$

RCPA adds the two scores before normalization:

$$\alpha_{ikj} = \mathrm{softmax}_j\left(\frac{s^r_{ikj} + s^c_{ikj}}{\sqrt{2d}}\right), \qquad y_{ik} = \sum_{j=1}^{N} \alpha_{ikj}\left(v^r_{ij} + v^c_{jk}\right). \tag{6}$$

Each candidate is weighted by its combined row and column score, and the same weight is applied to both values. The coupling occurs before softmax: a change in either branch can alter the weights of the other. By contrast, separate row and column attention computes two distributions and combines the outputs afterward. The resulting output remains a patch feature and can be passed directly to the next Transformer block. Normalization and positional encoding are detailed in Appendix B.1.

3.2 Balancing Sparse Connectivity and Global Coverage

Adapting PA to images requires limiting direct interactions without isolating distant patches. We therefore ask whether row column support permits global communication in two updates, and how sparse such support can be. The following propositions establish coverage and its degree requirement; joint normalization then specifies how the connected features interact.

Proposition 1 (two-hop coverage). Let $H, W \geq 2$ and let $\mathcal{N}(u)$ be the open neighborhood of u in $G_{H,W}$. For $u = (i, k)$ and $v = (p, q)$ with $i \neq p$ and $k \neq q$,

$$\mathcal{N}(u) \cap \mathcal{N}(v) = \{(i, q), (p, k)\}. \tag{7}$$

Thus, the graph has diameter two, with exactly two length-two paths for every nonadjacent pair.

This gives row column updates global structural reach after two layers: information from any patch can reach any other through a cross-corner location (Figure 2). Sparsity therefore restricts direct connections while preserving short communication paths. These graph intermediates describe connectivity; RCPA's shared index j instead pairs sources within one update.

Proposition 2 (degree required for two-hop coverage). Any graph on T vertices with diameter at most two and maximum degree Δ satisfies

$$T \leq 1 + \Delta + \Delta(\Delta - 1) = 1 + \Delta^2. \tag{8}$$

The rook's graph is regular with degree $H + W - 2$; on a square grid this is $2\sqrt{T} - 2$.

Two-hop coverage therefore requires $\Delta \geq \sqrt{T - 1}$, and square rook support attains this asymptotic order uniformly across patches. Its $O(\sqrt{T})$ neighbors per patch yield $O(T^{3/2})$ total interactions, balancing sparse local computation with global coverage and achieving an order-optimal maximum-degree result under the two-hop constraint.

From connectivity to joint computation. RCPA adds a paired-source update to this support. If p_j^r and p_j^c are branch softmax distributions with scale $1/\sqrt{2d}$, Equation 6 gives $\alpha_{ikj} = p_j^r p_j^c / \sum_\ell p_\ell^r p_\ell^c$. The normalized product makes both sources jointly determine each candidate's weight before either value is aggregated. RCPA thus retains PA's joint selection principle using grid-token states, reducing cubic all-pairs interactions to $O(T^{3/2})$ on square grids. Appendix D.1 accounts for channel projections and derives the grid sizes at which RCPA reduces the total attention-module MAC count.

3.3 Vision Transformer Floyd

We instantiate this operator in ViF using a standard pre-norm residual block:

$$u^{(\ell)} = x^{(\ell)} + \mathrm{DP}_{\ell,1}\Big(\mathrm{RCPA}\Big(\mathrm{LN}(x^{(\ell)})\Big)\Big), \tag{9}$$

$$x^{(\ell+1)} = u^{(\ell)} + \mathrm{DP}_{\ell,2}\Big(\mathrm{MLP}\Big(\mathrm{LN}(u^{(\ell)})\Big)\Big), \tag{10}$$

where DP is stochastic depth. Classification uses final LayerNorm, global average pooling, and a linear head. Appendix B describes the backbone organization and its adaptation to downstream tasks.

4 Experimental Results

Our experiments connect geometric relation learning to natural-image understanding and general visual recognition. On ClockBench, PA and RCPA Small outperform conventional backbones in relative-angle prediction under both IID and appearance or angular-distribution shifts. On VG150, ViF improves scene graph prediction across model scales, with consistent gains on the evaluated spatial predicates. ImageNet-1K, ADE20K, and COCO establish its utility beyond relation tasks, while component ablations explain the choices underlying RCPA.

4.1 Experimental Setup

We evaluate ViF at Tiny, Small, and Base scales, with widths of 192, 384, and 768. The ImageNet variants use 12, 16, and 16 blocks, respectively; ClockBench RCPA Small uses the 12-block variant. These models use p16 embeddings, MLP ratio four, RoPE, and query–key normalization.

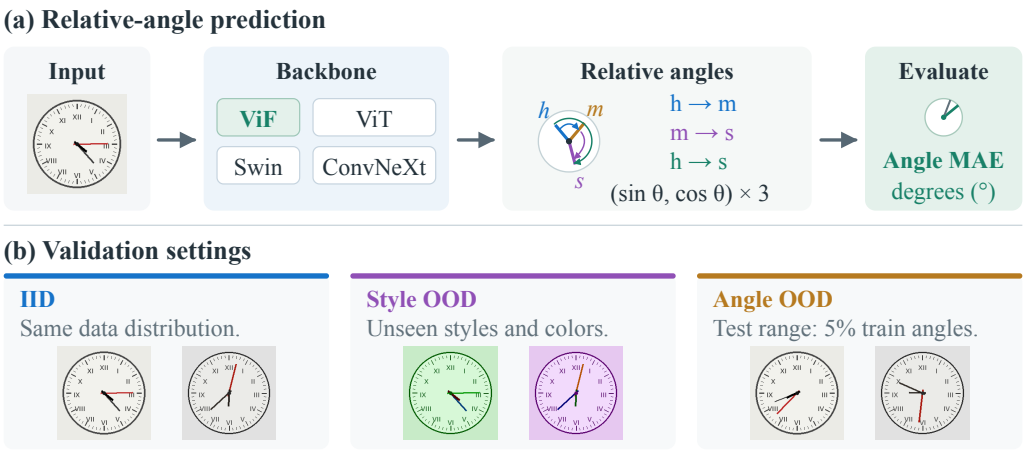


Figure 4: **ClockBench construction and evaluation.** (a) Predicting three directed relative angles from independently oriented clock hands, with sine/cosine regression and angular MAE evaluation. (b) Validation examples: IID preserves the training distribution, Style OOD changes appearance, and Angle OOD tests a range receiving 5% of training angles.

ViF ImageNet training uses 300 optimization plus 20 warm-up epochs with cosine learning-rate decay, and Muon optimizer. Augmentation includes RandAugment, random erasing, and scale-specific Mixup/CutMix settings. ADE20K uses UPerNet (Xiao et al., 2018) for 80k iterations with AdamW at learning rate 10^{-4}, layer-wise decay 0.8, and a polynomial schedule after 1,500 warm-up iterations; training uses random resizing, 512×512 crops, horizontal flips, and photometric distortion. COCO uses ViTDet-based detection (Li et al., 2022) for 25 epochs. VG150 uses 60k detector-pretraining iterations followed by 40k iterations with a Motif relation head (Zellers et al., 2018), keeping the pretrained detector frozen during relation training. The retrained ClockBench ViT Tiny/Small models and the paired Tiny RCPA comparison use AdamW at learning rate 2×10^{-4}, global batch 256, seed 42, and 100 total epochs including 10 warm-up epochs. Other ClockBench records use model-specific budgets and run lengths, detailed in Appendix C.

4.2 ClockBench: Synthetic Relation Modeling

We evaluate PA and RCPA on ClockBench to assess geometric relation learning and its generalization beyond the training distribution. The results show large gains over ViT, Swin, and ConvNeXt (Dosovitskiy et al., 2021; Liu et al., 2021b; 2022), demonstrating the effectiveness of PA-inspired computation for relative-angle prediction.

Benchmark curation. ClockBench renders independently oriented hands and annotates their three directed pairwise angles, isolating relative geometry from category recognition, as is shonw in Figure 4. IID evaluates the training distribution; Style OOD changes clock-face styles and colors; Angle OOD evaluates a half-circle receiving only 5% of training angles. Each setting has 20,000 training and 4,000 validation images, evaluated by angular MAE.

Learning relations across IID and distribution shifts. IID reveals both precise geometric modeling and efficient learning: PA leads within each scale, while RCPA lowers ViT's MAE from $7.69°$ to $2.23°$ at Tiny and from $5.99°$ to $1.96°$ at Small. On Style OOD, PA/RCPA outperform ViT at both scales; Tiny RCPA reaches $52.85°$ versus ViT's $82.20°$ by epoch 20. This early transfer supports rapid acquisition of relation features beyond color-specific appearance cues. On Angle OOD, PA is strongest ($4.68°/3.91°$), highlighting full pair-state modeling; RCPA sacrifices some accuracy but still beats all same-scale conventional baselines, reaching $23.04°$ versus ViT Small's $32.87°$.

Practical gains over Full PA. RCPA substantially reduces the computational cost of PA while retaining strong relation-modeling performance. At Small scale, it reduces computation by approximately $222\times$ and peak allocated memory by 89.1%, while increasing throughput by $7.44\times$ over

Table 1: **PA and RCPA outperform conventional baselines on visual relation learning, especially under appearance and angle-distribution shifts.** Training budgets and result-selection records differ across models; see Appendix C.

Model	Scale	IID	Style OOD	Angle OOD
patchsize=4				
ConvNeXt	Tiny	89.19	88.98	57.88
Swin	Tiny	17.22	64.38	59.31
patchsize=16				
ViT	Tiny	7.69	69.83	36.80
PA	Tiny	**1.32**	**21.76**	**4.68**
ViF (RCPA)	Tiny	2.23	29.55	32.21
ViT	Small	5.99	58.85	32.87
PA	Small	**0.98**	**17.23**	**3.91**
ViF (RCPA)	Small	1.96	26.50	23.04

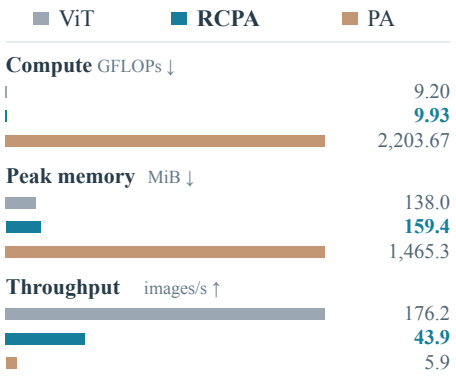


Figure 5: **ClockBench inference costs.** RCPA substantially reduces PA's computational complexity, making relation modeling over image patches feasible.

Table 2: **ViF improves scene graph detection and classification across model scales.** Gains over the corresponding ViT models span both overall and mean recall on VG150. Colored annotations report relative improvements; published methods are included as reference results.

	SGDet				SGCls
Backbone	R@50	R@100	mR@50	mR@100	R@100
Motifs (Zellers et al., 2018)	—	30.30	—	6.60	36.5
VCTree-HL (Tang et al., 2019)	—	31.30	—	8.00	38.8
BGNN (Li et al., 2021)	—	35.80	—	12.60	38.5
FCSGG (Liu et al., 2021a)	—	25.10	—	4.20	—
CoRF+T(+) (R50) (Adaimi et al., 2023)	—	18.10	—	4.10	28.7
CoRF+T(+) (Swin-S) (Adaimi et al., 2023)	—	20.00	—	4.50	30.6
ViT Tiny	28.80	32.66	6.04	7.23	36.2
ViF Tiny	**29.21** +1.42%	**33.26** +1.84%	**6.29** +4.14%	**7.58** +4.84%	**37.7** +4.14%
ViT Small	30.39	34.50	6.61	7.82	38.1
ViF Small	**32.27** +6.19%	**36.63** +6.17%	**6.89** +4.24%	**8.22** +5.12%	**39.0** +2.36%
ViT Base	31.78	35.81	7.00	8.20	39.2
ViF Base	**33.24** +4.59%	**37.47** +4.64%	**7.32** +4.57%	**8.57** +4.51%	**40.4** +3.08%

Full PA, shown in Figure 5. These savings bring its computation close to ViT's (9.93 versus 9.20 GFLOPs), while achieving lower relative-angle prediction error across IID, Style OOD, and Angle OOD settings. Together, these results show that RCPA preserves the benefits of PA-inspired relation modeling at a computational cost comparable to a conventional vision backbone.

4.3 VG150: REAL-IMAGE RELATION MODELING

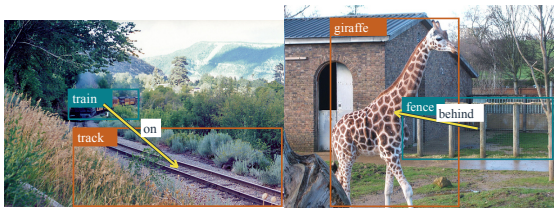
We evaluate ViF on VG150 to test relation modeling in natural scenes. VG150 contains 150 object categories and 50 predicates, representing images as subject–predicate–object triples that scene graph generation must recover (Krishna et al., 2017; Xu et al., 2017). Using a common detector-pretraining and relation-training pipeline, ViF improves both scene graph detection and classification across Tiny, Small, and Base scales. Detailed results are presented in Table 2.

Consistent gains across model scales. All four SGDet metrics improve at every scale, covering both overall recall and mean recall across predicates. The Small model increases R@100 by 6.17% over ViT, while Base reaches 37.47 R@100 and improves mR@100 by 4.51%. The accompanying SGCls gains extend the pattern to relation prediction with known object locations. Published systems provide additional context under their respective detector and training configurations.

Spatial relations in natural scenes. The predicate breakdown localizes a clear strength in spatial relationships: ViF specifically improves spatial predicates, including +2.90 and +2.98 points in per-predicate recall for *in front of* and *behind*. The gains persist in detection-conditioned recall,

Table 3: **ViF improves the spatial predicates specifically, while interaction predicates show mixed changes.**

Predicate	$\Delta mR@100$	ΔcR
Spatial		
in front of	+2.90	+3.21
behind	+2.98	+2.71
at	+1.67	+2.44
on	+2.30	+1.08
Interaction		
holding	+2.84	+1.76
carrying	+2.45	+1.53
sitting on	+0.91	+1.73
riding	−2.15	−4.47
eating	−0.04	−7.03
walking on	−3.33	−5.03



(a) Train on track (VG 2413708)

(b) Fence behind giraffe (VG 2338229)

Figure 6: **ViF recovers spatial relations missed by ViT despite successful object detection by both models.** The examples show (a) *train on track* and (b) *fence behind giraffe*.

Table 4: **ViF achieves better or comparable performance over ViT.** Results use 224×224 inputs.

Backbone	Params (M)	GFLOPs	Top-1 (%)	Top-5 (%)
ViT Tiny	5.72	2.51	74.25	92.01
ViF Tiny	6.57	2.54	**75.20**	**92.30**
Relative Δ (%)	–	–	+1.28	+0.32
ViT Small	22.05	9.20	**80.52**	**95.05**
ViF Small	33.80	13.20	80.49	95.02
Relative Δ (%)	–	–	-0.04	-0.03
ViT Base	86.57	35.13	81.52	94.74
ViF Base	133.67	52.29	**82.44**	**95.93**
Relative Δ (%)	–	–	+1.13	+1.26

Table 5: **ViF improves semantic segmentation at Small and Base scales.** Results report single-scale UPerNet performance on ADE20K.

Backbone	Params (M)	mIoU (%)	mAcc (%)
ViT Tiny	35.10	**39.03**	49.92
ViF Tiny	35.95	39.00	**50.31**
Relative Δ (%)	–	-0.08	+0.78
ViT Small	57.68	44.13	55.69
ViF Small	69.43	**46.48**	**58.73**
Relative Δ (%)	–	+5.33	+5.46
ViT Base	143.54	45.41	56.96
ViF Base	190.64	**47.62**	**59.49**
Relative Δ (%)	–	+4.87	+4.44

showing improved recovery of these relations among detected entities. Figure 6 makes the pattern concrete: ViF recovers *train on track* and *fence behind giraffe* when both models detect the entities but ViT misses the relation. Stronger entity detection also contributes to the end-to-end gain, raising relation coverage from 62.08 to 64.78.

Table 3 also shows declines on several interaction predicates, including riding, eating, and walking on. We hypothesize that these relations depend more strongly on action semantics and contextual knowledge, and therefore benefit less consistently from RCPA's spatial inductive bias. Explicit language-based semantic modeling may help connect visual configurations to these predicate meanings, suggesting a direction for extending the current vision-only model.

4.4 General Visual Transfer: ImageNet-1K, ADE20K, and COCO

We evaluate ImageNet-1K classification, ADE20K semantic segmentation, and COCO detection and instance segmentation to test RCPA as a general visual backbone (Deng et al., 2009; Zhou et al., 2017; Lin et al., 2014). Across these tasks, ViF extends its relation-oriented representations to category recognition and spatially dense prediction.

Visual recognition. ViF improves ImageNet top-1 accuracy by nearly one percentage point at both Tiny and Base scales, while matching ViT's Small-scale performance (Table 4). The Tiny model reaches 75.20% with 2.54 GFLOPs, compared with ViT's 74.25% at 2.51 GFLOPs. This gain at nearly identical computation shows that RCPA's relation-oriented update also supports effective category-level representations.

Dense visual prediction. Table 5 shows that the transfer gains are especially pronounced in semantic segmentation: ViF improves ADE20K mIoU by 2.35 and 2.21 points at Small and Base scales . While on COCO in Table 6, ViF Base reaches 48.50 box AP and 43.40 mask AP, improving

Table 6: **ViF Base improves both object detection and instance segmentation on COCO.** Box and mask AP each increase by 1.40 points over ViT Base, while Tiny and Small show smaller changes.

Model	Object detection ↑			Instance segmentation ↑			Img/s ↑	MiB ↓
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}		
ViT Tiny	43.8	**65.1**	47.5	39.0	**61.8**	41.5	8.9	817.4
ViF Tiny	**44.1**	65.0	**47.9**	**39.3**	**61.8**	**41.8**	9.5	831.1
Relative Δ (%)	+0.68	-0.15	+0.84	+0.77	0.00	+0.72	–	–
ViT Small	**47.3**	**68.5**	**51.2**	**41.9**	**65.4**	**44.7**	7.3	1405.9
ViF Small	47.0	67.2	50.8	41.6	64.8	44.6	7.7	944.8
Relative Δ (%)	-0.63	-1.90	-0.78	-0.72	-0.92	-0.22	–	–
ViT Base	47.10	67.40	50.70	42.00	65.00	44.70	4.4	2876.2
ViF Base	**48.50**	**69.50**	**52.60**	**43.40**	**66.80**	**46.50**	5.2	1224.9
Relative Δ (%)	+2.97	+3.12	+3.75	+3.33	+2.77	+4.03	–	–

both by approximately 3% relative to ViT. Together with the classification results, these improvements show that RCPA supports both image-level recognition and the spatial representations required for dense prediction.

4.5 Ablation Studies

The ablations support coupling row and column selection while keeping value aggregation simple. Under a common 100-epoch ImageNet-1K schedule for ViF Tiny, the default RCPA achieves the highest accuracy among the tested variants (Figure 7), supporting its shared query, joint weights, and additive values.

Joint source selection contributes to this performance by combining row and column evidence before aggregation. Replacing joint normalization with independent branch normalization on the same spatial support lowers top-1 accuracy by 1.348 points, from 72.694% to 71.346%. The two implementations use scales $1/\sqrt{2d}$ and $1/\sqrt{d}$, respectively, so this comparison evaluates their combined normalization and scaling choices.

Beyond joint selection, the results favor shared queries and simple additive value aggregation. Decoupling queries costs 2.782 points, while learned linear value fusion costs 4.072 points; additional gates also reduce accuracy. Together, these findings support using shared attention weights to select complementary spatial evidence and a direct sum to aggregate the selected values.

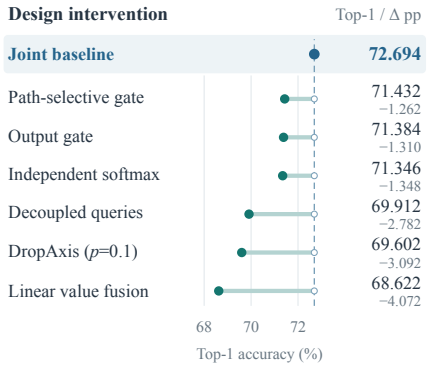


Figure 7: **The default RCPA design achieves the highest accuracy among the tested variants.** Joint normalization, shared queries, and additive value aggregation outperform the alternatives on ImageNet-1K.

5 Conclusion and Limitations

This work establishes PA-inspired relation updates as a promising direction for visual backbone design. RCPA makes this direction computationally feasible by exploiting the image grid, where sparse, uniform connections preserve global two-hop coverage with an asymptotically optimal maximum degree. The resulting ViF backbone supports geometric relation learning on ClockBench, improves VG150 scene graph detection, and transfers to classification and dense prediction.

Despite these results, broader relation benchmarks, including visual question answering, are needed to assess richer relational and multimodal tasks beyond relative-angle prediction and scene graphs. Further, although RCPA reduces spatial-interaction costs, channel projections and operator execution still contribute substantially to inference latency. Future work will optimize these projections, fuse attention kernels, and improve memory access to better translate the reduction in computation into end-to-end speedups, particularly at higher image resolutions.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing, literature search, figure preparation, and LaTeX formatting. We also used these tools to support code development, provide feedback on experimental design, and assist with the interpretation of results. The authors reviewed and revised the AI-assisted material, checked references against their sources, and verified reported results against experimental records. The authors take responsibility for the final manuscript, including its methods, results, claims, and supporting artifacts.

REPRODUCIBILITY STATEMENT

Appendix B specifies the backbone and task adapters; Appendix C describes ClockBench construction, evaluation, and training settings; Appendix D provides the computation analysis and attention-core measurement protocol. We plan to release code and configurations with the submission.

REFERENCES

George Adaimi, David Mizrahi, and Alexandre Alahi. Composite relationship fields with transformers for scene graph generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 52–64, 2023.

Awni Altabaa and John Lafferty. Disentangling and integrating relational and sensory information in transformer architectures. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Awni Altabaa, Taylor Webb, Jonathan D. Cohen, and John Lafferty. Abstractors and relational cross-attention: An inductive bias for explicit relational reasoning in transformers. In *International Conference on Learning Representations*, 2024.

Andries E. Brouwer and Willem H. Haemers. *Spectra of Graphs*. Universitext. Springer, New York, 2012.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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.

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.

Joshua E. Ducey, Jonathan Gerhard, and Noah Watson. The smith and critical groups of the square rook's graph and its complement. *arXiv preprint arXiv:1507.06583*, 2015.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-vocabulary 3d scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation*, pp. 5021–5028, 2024.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers, 2019.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, 2018.

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.

Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. EGTR: Extracting graph from transformer for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24229–24238, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. LLM4SGG: Large language models for weakly supervised scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28306–28316, 2024a.

Manjin Kim, Paul Hongsuck Seo, Cordelia Schmid, and Minsu Cho. Learning correlation structures for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18941–18951, 2024b.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11109–11119, 2021.

Rongjie Li, Songyang Zhang, and Xuming He. SGTR+: End-to-end scene graph generation with transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28076–28086, 2024b.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision – ECCV 2022*, pp. 280–296. Springer, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755, 2014.

Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11546–11556, June 2021a.

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*, pp. 10012–10022, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, and Bin Dai. DriveWorld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15522–15533, 2024.

Shanka Subhra Mondal, Jonathan D. Cohen, and Taylor Whittington Webb. Slot abstractors: Toward scalable abstract visual reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36088–36105, 2024.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring spatial representations for vision-language-action models. In *Robotics: Science and Systems*, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Atharva Sehgal, Arya Grayeli, Jennifer Sun, and Swarat Chaudhuri. Neurosymbolic grounding for compositional world models. In *International Conference on Learning Representations*, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6619–6628, 2019.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *Computer Vision – ECCV 2022*, volume 13684 of *Lecture Notes in Computer Science*, pp. 459–479. Springer, 2022.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Computer Vision – ECCV 2020*, pp. 108–126, 2020.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5419, 2017.

Jingcheng Yu, Mingliang Zeng, and Qiwei Ye. FloydNet: A learning paradigm for global relational reasoning, 2026.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5831–5840, 2018.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3d vision-language-action generative world model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61229–61245. PMLR, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025.

A Related Work

RCPA connects three lines of research: visual relation modeling, explicit relational representations, and structured spatial attention. Together, these lines motivate a backbone that combines relation-oriented updates with an interaction pattern suited to image grids.

Visual relations. Visual relation models show how interactions between entities can support understanding beyond individual object recognition. Relation Networks and object-relation modules combine information from pairs of objects or regions (Santoro et al., 2017; Hu et al., 2018). Building on this idea, scene graph methods predict subject–predicate–object structures through message passing, scene context, or transformer decoders (Xu et al., 2017; Zellers et al., 2018; Li et al., 2024a); EGTR extracts relation information from a detector's self-attention (Im et al., 2024). These approaches organize relational computation around objects and task outputs. RCPA brings a related motivation to the spatial feature mixer, allowing the resulting backbone features to serve several visual tasks.

Relational representations. Explicit relational representations provide a way to organize computation around relationships among inputs. Abstractors separate relational information from object attributes using relational cross-attention (Altabaa et al., 2024). Slot Abstractors extend this approach to object-centric visual representations (Mondal et al., 2024), while Dual Attention Transformers process sensory and relational information with separate attention mechanisms (Altabaa & Lafferty, 2025). Pair-based architectures further represent relationships as states that can be updated through a shared intermediate: Evoformer uses triangle operations (Jumper et al., 2021), and FloydNet's PA constructs candidate keys and values from two pairs sharing an intermediate element (Yu et al., 2026). The latter construction motivates our joint row–column update, which adapts shared-intermediate composition to the image grid while avoiding all-pairs storage.

Structured visual attention. The spatial organization of images offers interaction patterns that reduce the cost of global feature mixing. Non-local networks and ViT aggregate features across spatial positions (Wang et al., 2018; Dosovitskiy et al., 2021); axial attention, CCNet, CSWin, and MaxViT organize this mixing through axes, criss-cross neighborhoods, stripes, and block/grid patterns, respectively (Ho et al., 2019; Wang et al., 2020; Huang et al., 2019; Dong et al., 2022; Tu et al., 2022). StructViT additionally learns spatial correlation structure within the backbone (Kim et al., 2024b). RCPA uses image structure to organize paired row and column candidates, then scores them jointly before normalization. This joint selection connects the efficiency of structured attention with the shared-intermediate update motivating our design.

B Architecture and Implementation Details

ViF integrates RCPA into a Transformer backbone whose spatial features support both classification and dense prediction. The architecture separates the shared feature update from the task-specific readout: RCPA mixes spatial information within the backbone, and lightweight adapters expose the resulting features to each prediction head.

B.1 Backbone organization and attention update

The ViF variants share a common block structure, so model scale changes preserve the same form of spatial interaction. A convolutional 16×16 patch projection constructs the initial grid; pre-norm residual blocks then alternate RCPA with an MLP of expansion ratio four. Stochastic depth increases linearly across blocks.

Within each attention update, positional encoding and normalization prepare the features for joint row–column selection. The ImageNet and downstream models, as well as the reported ClockBench RCPA model, use RoPE with additive absolute positional embeddings disabled. Queries and both path-specific keys receive RMS normalization. On square grids, the two sets of logits are added before a single softmax over the common coordinate index, and the corresponding values are summed. Thus, the two spatial sources participate in a shared selection step.

B.2 Adaptation to downstream tasks

Task-specific readouts connect the shared backbone to the spatial granularity of each prediction problem. Classification first normalizes and averages the final grid, then applies a linear classifier. Dense prediction instead exposes spatial feature maps at multiple scales, preserving their locations for detection and segmentation.

For COCO, the ViTDet adapter converts the final normalized grid to channel-first form and constructs four pyramid levels for the detection neck. Visual Genome uses the corresponding ViF–FPN or ViT–FPN adapters. For ADE20K, intermediate backbone grids feed a MultiLevelNeck and UPerNet; the selected zero-based block indices are $(2, 5, 8, 11)$ for 12-layer models and $(3, 7, 11, 15)$ for 16-layer models. These interfaces allow the same token mixer to support image-level and spatially dense outputs.

C ClockBench

ClockBench examines relative-angle prediction under changes to visual appearance and geometry. The additional analyses connect three observations: RCPA achieves lower error early in a shared-budget Tiny comparison, PA and RCPA preserve angle-dependent representations across paired color changes, and geometric deformation exposes a distinct generalization challenge. We first define the task and evaluation, then develop each observation in turn.

C.1 Task design and evaluation

Separating angle relations from their visual rendering allows ClockBench to test how predictions respond to different forms of variation. Each image contains hour, minute, and second hands with independently sampled angles (h, m, s), and the targets are the three directed clockwise differences:

$$r_{hm} = (m - h) \bmod 360, \qquad r_{ms} = (s - m) \bmod 360, \qquad r_{hs} = (s - h) \bmod 360. \tag{11}$$

Each target is encoded as sine and cosine, yielding six outputs trained with mean-squared error. At evaluation, `atan2` decodes each pair into an angle. The circular error

$$e(\hat{r}, r) = \min(|\hat{r} - r|, 360 - |\hat{r} - r|) \tag{12}$$

is averaged over the three relations and all validation examples to obtain MAE.

The splits vary appearance and angle coverage separately. IID samples all angles from $[0, 360)$ in both partitions. Style OOD preserves this support while using disjoint clock-face styles and palettes, testing transfer across appearance changes. Angle OOD concentrates validation on $[180, 360)$, which receives 5% of training draws; the remaining 95% lie in $[0, 180)$. It therefore tests generalization to a sparsely represented angular range with overlapping training support. Geometric deformation further changes the rendered shape, as discussed below.

All images are 224×224 RGB renderings with randomized hand widths and lengths and predefined face/color palettes. Each split contains 20,000 training and 4,000 validation images, generated with seeds 42 and 10042, respectively.

C.2 Learning under a shared training budget

RCPA achieves lower prediction error than ViT at the same early training budget in the Tiny comparison. At epoch 20, corresponding to 1,580 optimizer steps, IID MAE is 6.5554° for RCPA and 45.2033° for ViT. The difference also appears on Style OOD, where the respective errors are 52.8463° and 82.2025°. These fixed-step results complement the final accuracies in Table 1 by showing that the RCPA model learns useful angle predictions earlier in this paired experiment.

To compare the models at a common budget, both use the same V3 datasets, seed 42, AdamW with learning rate 2×10^{-4}, weight decay 0.02, clipping at 1, batch 256, 10 warm-up epochs, and 100 total epochs. The comparison concerns the Tiny ViT/RCPA pair; the PA results are assessed through task accuracy and the representation analysis below.

C.3 Representation stability under appearance changes

PA and RCPA preserve angle-dependent representation structure across paired color changes. In Figure 8(a), all three architectures exhibit circular similarity patterns on IID images, indicating that their representations organize examples by relative angle. This shared IID pattern motivates a more discriminating comparison: holding hand geometry fixed while changing clock-face and hand colors.

Under that intervention, Figure 8(b) shows that the circular structure remains clearer for PA and RCPA than for ViT. Together with the reported color-induced prediction changes, the cross-color similarity patterns connect angle organization to stability under appearance variation. The analysis describes the learned representations; relative-angle prediction can also be implemented by estimating individual orientations and subtracting them, so this observation leaves the internal computational mechanism open.

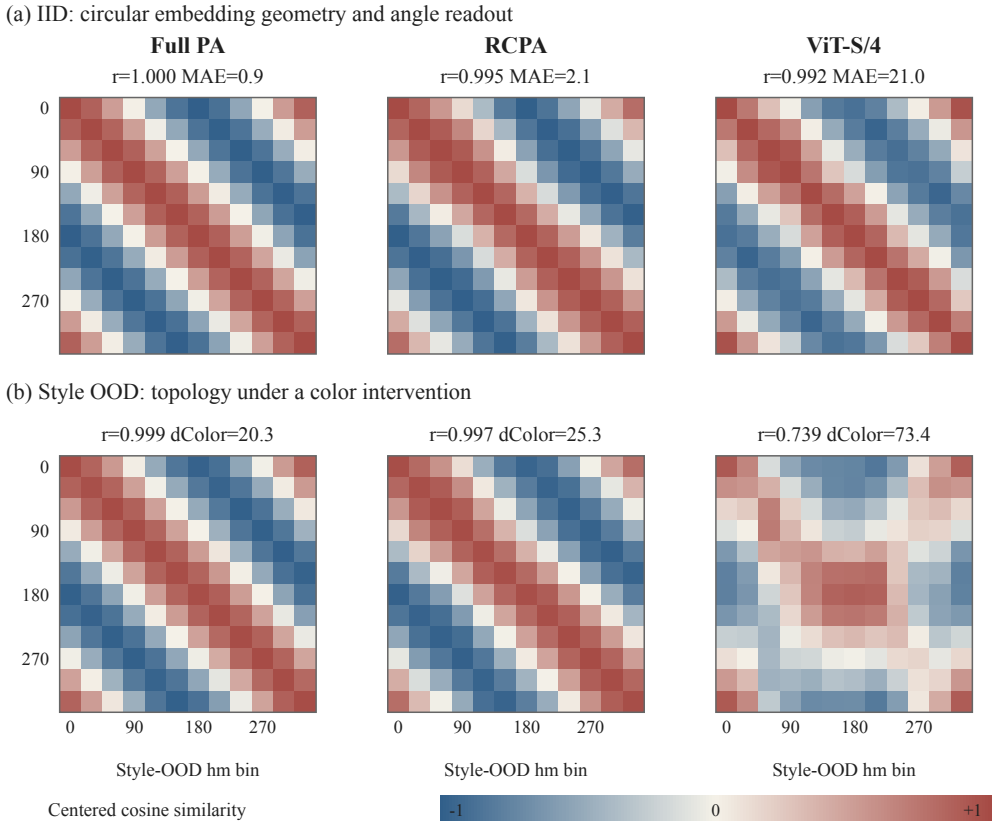


Figure 8: **PA and RCPA preserve angle-dependent representation geometry across color changes.** Centered cosine similarity between directed-angle bin centroids: (a) within IID images and (b) across paired images with identical hand geometry but different colors. Colors encode similarity; r is correlation with an ideal circular-angle kernel, and dColor is the mean circular prediction change in degrees.

The figure uses one retained style-run checkpoint per architecture. Its readout MAEs characterize those checkpoints, while Table 1 reports best-validation results.

C.4 Generalization under geometric deformation

Robustness to appearance changes does not extend uniformly to geometric deformation. Table 7 shows substantial errors across the tested models: PA Tiny obtains $53.70°$, while Swin Tiny obtains the lowest error, $47.70°$. Thus, strong IID angle prediction and stable color-conditioned representations address only part of the generalization problem.

The distinction follows from the intervention itself. Color changes preserve the rendered hand geometry, whereas squeezing, stretching, or shearing changes it. Here, these transformations are applied only at validation time and retain the original stored angle labels. Consequently, the experiment measures prediction stability relative to those labels under image deformation, extending the appearance analysis to a different form of distribution shift.

Table 7: **Geometric deformation remains challenging across architectures.** Angular MAE (degrees, lower is better) after validation-only squeeze, stretch, or shear with the original angle labels retained.

Model	Deformation MAE
PA Tiny	53.70
ViT Small	81.60
ConvNeXt Tiny	88.90
Swin Tiny	47.70

The PA result was checked against the original evaluation output; the other results are retained from archived summaries. RCPA deformation results are unavailable.

Training and result selection. The current ViT Tiny/Small rows and the paired Tiny RCPA experiment use the batch-256, seed-42 recipe in Section 4.1, with 100 total epochs including 10 warm-up epochs. The ViT Small row replaces the earlier batch-2048 run. Historical PA Tiny uses batch 32 for IID/Style OOD and batch 16 for Angle OOD; its IID result is the best observed before interruption near epoch 55, while the other splits completed 100 epochs. RCPA Small uses batch 256, whereas Swin and ConvNeXt use batch 2048. These records therefore share task and evaluation settings but differ in optimization budgets.

ConvNeXt's 89.19° IID MAE is close to the 90° expectation of a constant predictor under uniform angles, indicating little useful angle prediction in that run. It is a result of the recorded training recipe; convergence and tuning adequacy have not been established from the available evidence.

D Finite-size cost accounting and attention-core diagnostic

RCPA's computational benefit grows with the spatial grid: sparse row–column interactions reduce the growth of spatial-mixing costs, while additional channel projections impose a cost at finite resolutions. We first derive the point at which interaction savings outweigh these projections, then examine how the attention core's scaling appears in measured latency and memory. Together, the analyses explain why small-grid operation costs and large-grid efficiency can favor different methods.

D.1 Finite-size computation costs

Channel projections can dominate RCPA's attention cost at ordinary image resolutions despite its lower spatial-interaction complexity. To separate these contributions, let $T = s^2$ denote the number of spatial tokens and C the channel width. SA has three input projections and one output projection; RCPA has five input projections (Q, K_r, K_c, V_r, V_c) and one output projection, with the two rotated queries sharing the learned query projection. The dominant matrix costs are therefore

$$M_{\mathrm{SA}} = 4TC^2 + 2T^2C, \tag{13}$$

$$M_{\mathrm{RCPA}} = 6TC^2 + 4T^{3/2}C. \tag{14}$$

The first term accounts for channel projections and the second for spatial mixing. One MAC corresponds to two matrix-product FLOPs; these expressions exclude bias, normalization, RoPE, softmax, and other elementwise operations.

The balance shifts toward RCPA as the grid grows. For equal width and depth, its matrix cost becomes smaller when

$$T - 2\sqrt{T} > C, \qquad s > 1 + \sqrt{1 + C}. \tag{15}$$

For widths $192, 384, 768$, the smallest integer grid sides satisfying this condition are 15, 21, and 29. A 224×224 image with patch size 16 yields a 14×14 grid, below all three thresholds. Table 8 makes the projection contribution explicit.

Table 8: **Attention-module computation at a 14×14 patch grid.** Projection and attention-core costs are reported in millions of MACs for feature width C. The total ratio is RCPA/SA, including both projections and the attention core.

C	SA proj./core (M MACs)	RCPA proj./core (M MACs)	Total ratio
192	28.90 / 14.75	43.35 / 2.11	1.041
384	115.61 / 29.50	173.41 / 4.21	1.224
768	462.42 / 59.01	693.63 / 8.43	1.346

At Base width, projections account for 98.8% of RCPA's module MACs. Increasing the token count therefore provides a direct route to realizing the interaction savings: at fixed 224px resolution, patch size 8 gives $T = 784$, still below the Base crossover, whereas patch size 4 gives $T = 3136$, above it.

Block depth and the MLP further determine how this module-level saving translates to a backbone. An MLP ratio of four adds $8TC^2$ per block to both methods. In addition, the Small and Base ViF classifiers have 16 blocks, compared with 12 in the corresponding DeiT references; their block-stack MAC ratios at $T = 196$ are approximately 1.45 and 1.50. For depth ratio $\rho = L_{\mathrm{RCPA}}/L_{\mathrm{SA}}$, the crossover becomes $T - 2\rho\sqrt{T} > (7\rho - 6)C$, giving minimum grid sides 38/52 for Small/Base when $\rho = 4/3$. These thresholds describe matrix computation in the compared block stacks; complete task costs additionally depend on patch embedding, class tokens, prediction heads, and any window/global-attention mixture.

D.2 Runtime and memory scaling

The measured RCPA attention core achieves both lower latency and lower peak memory on large token grids. At width $C = 768$, its speed advantage grows from $6.27\times$ on a 56×56 grid to $19.80\times$ on a 112×112 grid. The latter case also reduces peak allocated memory from 7267.2 to 302.1 MiB. These results show how sparse spatial mixing benefits the explicit attention implementation as the token count increases.

To isolate this contribution, we remove channel projections from both methods, yielding the core MAC ratio $M_{\mathrm{RCPA,core}}/M_{\mathrm{SA,core}} = 2/\sqrt{T}$. The comparison includes score computation, scaling, softmax, value aggregation, and head merging. Table 9 shows the transition from small-grid overhead to large-grid savings at $C = 768$.

Table 9: **Attention-core latency and peak allocated memory across grid sizes.** Grid denotes the side length of a square patch grid. Measurements use feature width $C = 768$, batch size 1, and FP16 inputs on a V100. SA/RCPA is the latency ratio; memory is reported in MiB. Both methods exclude channel projections.

Grid	SA ms	RCPA ms	SA/RCPA	SA MiB	RCPA MiB
14	0.103	0.383	0.27	10.7	11.2
28	0.153	0.306	0.50	39.7	21.5
56	2.301	0.367	6.27	473.9	66.1
112	28.452	1.437	19.80	7267.2	302.1

At 14×14, RCPA is slower even though its core uses only one seventh of SA's MACs. Because both cores exclude projections, this gap shows that matrix-operation counts alone do not determine latency. Kernel-launch overhead and the efficiency of smaller matrix operations are possible contributors. As the grid grows, the reduced interaction count is accompanied by a substantial measured speed advantage. The memory trend is likewise consistent with materialized attention storage growing as $O(hT^{3/2})$ for square-grid RCPA and $O(hT^2)$ for SA, where h is the number of heads. Measured peak memory also includes inputs and other live buffers.

The two analyses therefore identify complementary requirements for efficient RCPA inference: sufficiently large grids expose its spatial-mixing advantage, while efficient projection and operator execution remain important for the complete model. The measurements here characterize the explicit attention core; the main task tables report full-model costs.

Measurement protocol. SA and RCPA v1 use contiguous, precomputed head-major FP16 inputs on a V100, with batch 1 and head dimension 64. The core excludes projections, QK normalization, RoPE, and MLP; Flash/SDPA, compilation, and TF32 are disabled. Widths $192/384/768/1024$ and grid sides $14/28/56/112$ are measured. Reported latency is the median of five 50-iteration repeats after 30 warmups for large grids, and seven 250-iteration repeats after 150 warmups for grids 14/28. Large-grid Base CVs are below 0.5%; the 14-grid RCPA Base CV is 13.2%. Memory is total peak allocated memory, including precomputed inputs and warmed backend allocations.