

Peer Review

Review of: "Translution: Unifying Self-Attention and Convolution for Adaptive and Relative Modelling"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes Translution, an operation in which the query, key, and value projections are no longer shared across positions. Instead, each relative offset (δ_x, δ_y) between two patches gets its own matrices W_{δ_x, δ_y}^q , W_{δ_x, δ_y}^k , W_{δ_x, δ_y}^v (Eq. 1, §3.3). The authors frame this as a unification: convolution becomes the case where attention is a fixed 0/1 mask over the kernel, and self-attention becomes the case where every offset shares one projection. Because the full version needs $(2H-1) \times (2W-1)$ matrices per projection (729 for ViT/16 at 224×224 , §3.4), they also introduce α -Translution, which routes the offset-specific part through small bottlenecks $C_1 = C_2 = 8$ and adds it to the ordinary self-attention path (Eq. 2). A 1D causal version for text is given in Appendix C, and a memory-saving rewrite of the value sum in Appendix D.

Empirically, the claim is that both variants beat self-attention with absolute positional embedding at equal architecture: on a synthetic dynamic MNIST (Table 2), on ImageNet-1K trained from scratch with small ViTs (Table 3), against six existing positional schemes (Table 4), and on OpenWebText perplexity with short sequences of 160 tokens (Table 7). An ablation replaces relative matrices with per-pair absolute matrices to argue that the gain is not just extra parameters (Table 5).

2. Whether the conclusion holds on the data

I like the core idea a lot. It is simple to state, it makes a real design choice explicit (shared versus offset-specific projections), and the authors are candid that the full operation does not fit on current hardware and that their validation is "preliminary" (§5). The dynamic MNIST setup is a clean probe for the

property they care about. Seeing the gap open from Static→Static to Dynamic→Dynamic is persuasive: in Table 2, ViT-A/12 self-attention drops to 92.64 while α -Translution holds 97.31. The same pattern appears at ViT-A/7 (93.90 against 98.57). Appendix F then adds absolute embeddings back into Translution and shows Static→Dynamic falling from 36.24 to 16.50 (Table 9, Translution rows). That is a nice, direct piece of evidence for the paper's argument about absolute positions.

I also checked whether the parameter counts for full Translution follow from Eq. 1. For ViT-A/12 on 84×84 input, the grid is 7×7 , so there are $13 \times 13 = 169$ offsets. Three projections of 192×192 over six layers give $3 \times 169 \times 36,864 \times 6 = 112,140,288$, about 112.1M. Taking the 2.7M self-attention model, removing its shared projections ($3 \times 36,864 \times 6 \approx 0.66\text{M}$), and adding the relative CLS weights of Appendix H (three directions $\times$ three projections $\times 36,864 \times 6 \approx 1.99\text{M}$) gives $2.7 - 0.66 + 112.14 + 1.99 \approx 116.2\text{M}$, which matches Table 2. So the headline counts look internally consistent, which I appreciate.

Where I think the conclusions run ahead of the data is in three places.

Parameter and compute matching. Table 5 answers one version of the parameter question (absolute per-pair matrices do worse), but only on MNIST. On ImageNet, the gains come with large parameter increases, and the paper's own Table 3 lets a reader compare across architectures. ViT-B/16 self-attention (5.7M) reaches 73.51 Top-1, while ViT-A/16 α -Translution (10.7M) reaches 69.28. ViT-C/56 self-attention (25.3M, 64.15) beats ViT-A/56 Translution (38.5M, 52.41). None of this contradicts the equal-architecture claim. It does mean the abstract's phrase about "the next generation of deep neural networks" needs a per-parameter or per-FLOP comparison before it can be read as practical advice. FLOPs, throughput, and peak memory are not reported for any row.

What α -Translution actually gains from. Table 10 is revealing. With the shared W^q , W^k , W^v removed, the purely relative model at $C_1 = C_2 = 8$ gets 41.81 Top-1 (row " $C_1=C_2=8$, $\times$ "), which is below self-attention without any positional embedding (42.49, row " $C_1=C_2=0$ ") and well below self-attention with positional embedding (46.28, Table 4). The benefit, therefore, comes from adding a low-rank offset-specific term on top of standard attention. That is still a worthwhile finding. I would state it this way in §3.4 rather than presenting α -Translution as a lighter copy of Translution.

Effect sizes and noise. Every number is a single run. In Table 4, the existing relative schemes sit between 46.35 and 46.65 Top-1, a spread of 0.30 points, which I would not try to rank without seeds. The α -Translution gain over self-attention ($46.28 \rightarrow 48.36$, i.e., +2.08) and the Translution gain ($\rightarrow 52.41$) are larger and probably real, but a mean and standard deviation over three seeds for at least the ViT-A/56

rows would settle it cheaply, since these are the smallest models. On text, the improvement shrinks with scale: Table 7 shows $60.40 \rightarrow 57.97$ for GPT-A-160, $54.82 \rightarrow 52.72$ for GPT-B-160, but only $39.88 \rightarrow 39.25$ for GPT-C-160. A 0.63 perplexity gap at the largest size is exactly where variance matters.

The ImageNet regime deserves one more sentence. ViT-A/56 sees a 4×4 grid of 16 patches, and self-attention reaches 46.28 Top-1. That is far from standard ViT operating points, and relative encoding plausibly helps most when positional information is scarce. The ViT-A/16 and ViT-B/16 rows (+4.57 and +2.69 for α -Translution) are the most relevant evidence here, and I would give them more weight in the text.

I also found a few factual inconsistencies that should be corrected (listed with coordinates in section 4): the Translution Static $\rightarrow$ Dynamic score differs between Table 2 and Table 5, one Table 3 row label appears to be wrong, and the text points to "Section D" for a comparison that lives in Appendix E.

3. Reproducibility

The code link is given in the abstract, which is a good start, and Appendix I lists the optimiser, learning rate, weight decay, epochs, batch size, and the ImageNet augmentation pipeline (RandAugment 2/9, CutMix $\alpha = 1.0$, MixUp $\alpha = 0.8$). That is more than many architecture papers provide.

Some items I could not find in the Qeios text:

- Random seeds, number of runs, and any variance estimate.
- How dynamic MNIST is generated: number of training and test images, whether positions are sampled once per image or per epoch, and whether a fixed seed produces the released split. §4.1.1 and Figure 4 describe it only qualitatively.
- Hardware, GPU-hours, and wall-clock time per run. §4 and §5 speak only of "limited computational resources," without naming the GPUs or the time each experiment took.
- The 2–3 TB single-GPU figure in §4 and §5. I could not reconstruct it from the text; a short derivation (parameters, optimiser state, activations of the $N \times N \times C'$ value tensor) would make it checkable.
- A per-component parameter breakdown for α -Translution. From Eq. 2, I expected the extra cost for ViT-A/16 to be roughly $729 \text{ offsets} \times 8 \times 8 \times 3 \text{ projections} \times 6 \text{ layers}$, about 0.84M, yet Table 3 shows $3.0\text{M} \rightarrow 10.7\text{M}$. The difference may come from per-head matrices or the CLS weights, but I could not tell from the text.
- The OpenWebText schedule. Appendix I.3 says the cosine schedule runs "over every 3.2k iterations" across 9.6 million iterations, which reads like warm restarts; please confirm. $9.6\text{M} \times 8 \times 160 = 12.29\text{B}$

tokens against 9B training tokens (§4.2), so about 1.37 passes. Whether all three GPT sizes used the same budget is not stated.

4. What specifically to fix

1. **Table 2 vs Table 5 and Table 9.** Translution (ViT-A/12, relative) Static→Dynamic is 36.40 in Table 2 but 36.24 in Table 5 and Table 9. Only one can be the reported run.
2. **Table 3, row after "Translution runs out of memory".** A second "ViT-A/32" appears with 22.9M parameters and 73.62 Top-1, while ViT-A/32 self-attention is already listed at 3.5M and 57.63. I assume this is ViT-C/32; please check. The caption also says "patch sizes of 56 and 32" although the table contains ViT-A/16 and ViT-B/16.
3. Add seeds. Three runs for the ViT-A/56 rows of Table 4 and the GPT-C-160 rows of Table 7 would be enough to show which differences survive noise.
4. **§4.1.2 and Table 3.** I would add FLOPs, throughput, and peak training memory next to #Params, and at least one parameter-matched self-attention baseline (for example, a wider ViT-A/16 at around 10.7M) so the reader can see the trade-off directly.
5. **§3.4 and Appendix G.** State plainly that the offset-specific branch alone at $C_1 = C_2 = 8$ underperforms (Table 10: 41.81 against 42.49 for $C_1 = C_2 = 0$), and that α -Translution works as an additive low-rank relative term on top of standard attention.
6. On dynamic MNIST, a small CNN baseline would help. Convolution is the other half of the claimed unification and is translation-equivariant by construction, so its Static→Dynamic number is the natural reference point for the 34.90 and 36.24 of Translution.
7. **Cross-references.** §2 (end of the first paragraph) and §4.1.2 refer to "Section D" for the formal comparison of position schemes; that material is Appendix E. Appendix D is the memory-efficient implementation.
8. **Notation.** In the convolution case of the unification (§3.3), the mask condition reads $(\delta_i, \delta_j) \in \text{kernel}$; it should be (δ_x, δ_y) . In the §4.1.3 ablation, the indices are written $i \in [1, H \times W]$ and $y \in [1, H \times W]$; the second should be j .
9. Small naming slips: "Translation with ViT-A/7" in the Table 5 caption, " α -Transformer" at the start of Appendix D, and in the Table 10 caption.
10. **Reference for OpenWebText.** §4.2 cites reference [30], which in the list is Gao et al., "The Pile". OpenWebText is a different corpus and should get its own citation.

11. Soften the last sentence of the abstract. "Superior accuracy compared to self-attention" at equal architecture is supported; "the next generation of deep neural networks" would need the scale and efficiency evidence that §5 itself says is still missing.
12. **Causal 1D Translution, Appendix C.** The relative value case " $\forall$, otherwise" is hard to parse as rendered on Qeios. I would write explicitly that $v_{i,j}$ is unused (or zero) for $j > i$.

5. Recommendation

Minor revisions. The idea is clear and original enough to be worth building on, the experiments cover vision and language, and several of them (dynamic MNIST, Tables 5, 9, and 10) are well designed to test the specific mechanism. The fixes I am asking for are mostly bookkeeping: correct two table inconsistencies and a few references, add seeds, and report compute next to parameters. With those in place, I think the paper's main claim at equal architecture will stand on firmer ground, and the broader claims in the abstract can be matched to what the evidence actually shows.

Declarations

Potential competing interests: No potential competing interests to declare.