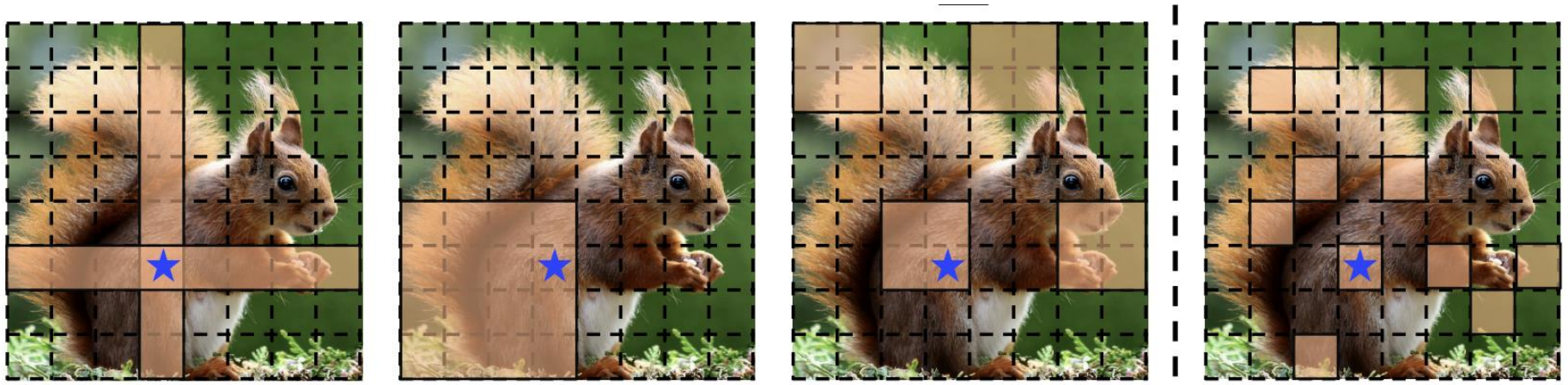
# Random Wins All: Rethinking Grouping Strategies for Vision Tokens

Qihang Fan

Institute of Automation, Chinese Academy of Sciences

# [CVPR2026] *Random Wins All: Rethinking Grouping Strategis for Vision Tokens.*

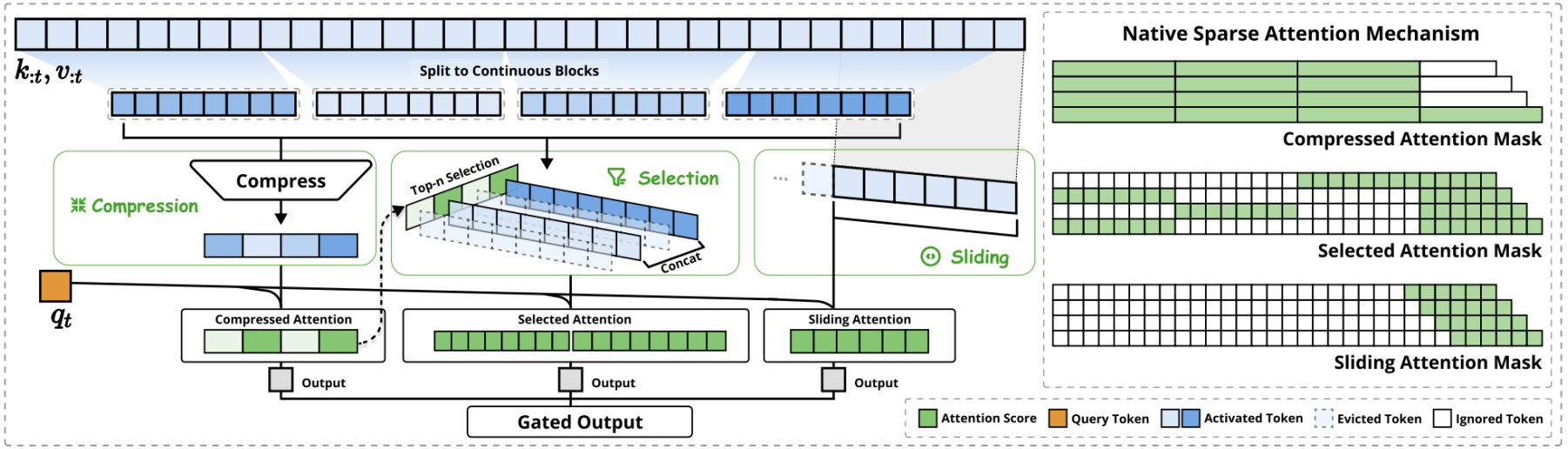
- **TL;DR:** Explore the necessity of specially designed sparse attention mechanisms.



(a) Cross-Window Grouping (b) Window Grouping (c) Bi-level Routing Grouping (d) Random Grouping

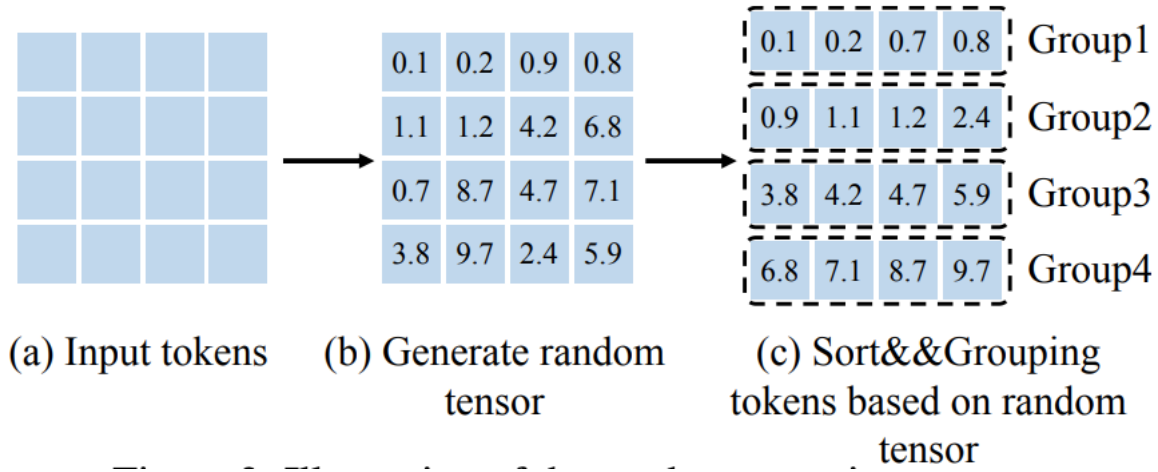
# [CVPR2026] *Random Wins All: Rethinking Grouping Strategis for Vision Tokens.*

- **Motivation:**
  - To model the relationships between tokens in long sequences, many specially designed sparse attention mechanisms have been proposed.
  - The vast majority still require specialized hardware optimization.
  - Are these specially designed approaches necessary?



# [CVPR2026] *Random Wins All: Rethinking Grouping Strategis for Vision Tokens.*

- **Method:**
  - A simple random grouping strategy.
  - Simple && fast && general



# [CVPR2026] *Random Wins All: Rethinking Grouping Strategis for Vision Tokens.*

- **Method:**
  - Intuitively, this method might not seem effective, but in practice, it brings significant improvements.

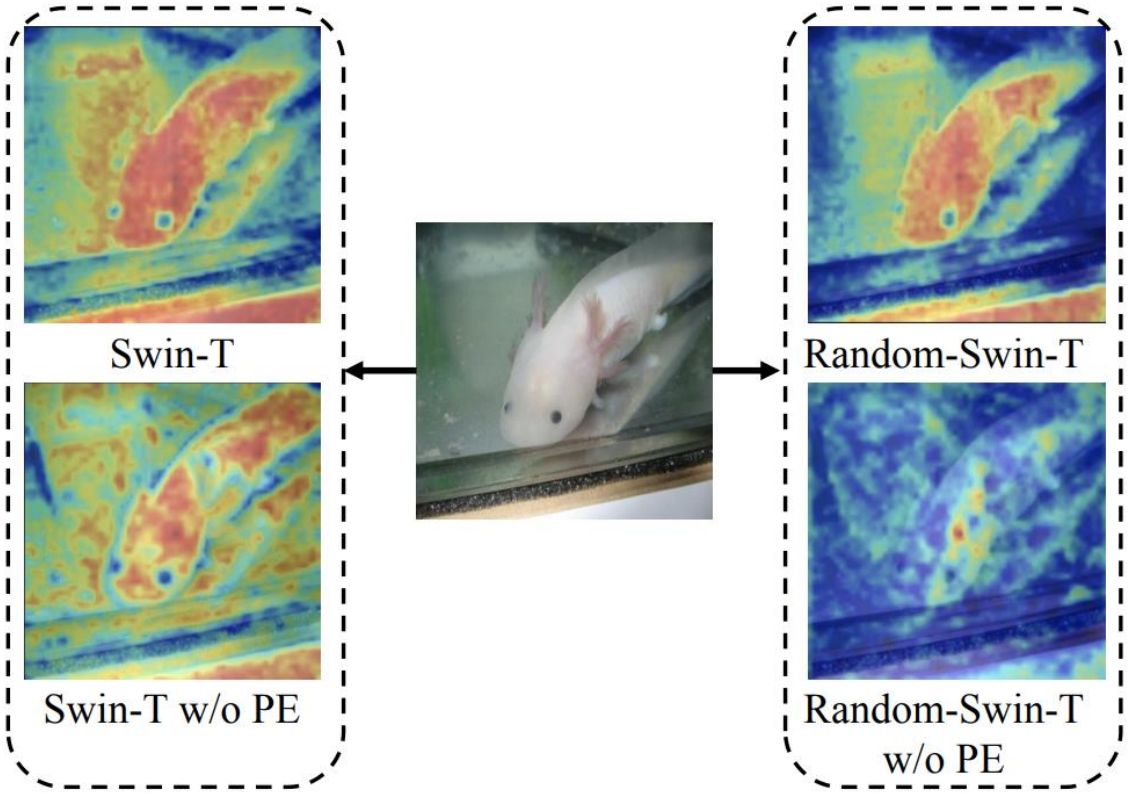
| Backbone | Params (M) | FLOPs (G) | Mask R-CNN 1× $AP^b$ | $AP^b_{50}$ | $AP^b_{75}$ | $AP^m$ | $AP^m_{50}$ | $AP^m_{75}$ | Params (M) | FLOPs (G) | RetinaNet 1× $AP^b$ | $AP^b_{50}$ | $AP^b_{75}$ | $AP^b_S$ | $AP^b_M$ | $AP^b_L$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Partition-Based Backbone | | | | | | | | | | | | | | | | |
| Swin-T [23] | 48 | 267 | 43.7 | 66.6 | 47.7 | 39.8 | 63.3 | 42.7 | 38 | 248 | 41.7 | 63.1 | 44.3 | 27.0 | 45.3 | 54.7 |
| Random-Swin-T | 48 | 267 | **46.0** | **68.1** | **50.5** | **41.9** | **65.3** | **45.4** | 38 | 248 | **44.3** | **65.8** | **46.9** | **29.7** | **47.9** | **57.8** |
| Swin-S [23] | 69 | 359 | 45.7 | 67.9 | 50.4 | 41.1 | 64.9 | 44.2 | 60 | 339 | 44.5 | 66.1 | 47.4 | 29.8 | 48.5 | 59.1 |
| Random-Swin-S | 69 | 359 | **48.0** | **69.5** | **51.9** | **43.2** | **67.1** | **47.0** | 60 | 339 | **46.6** | **67.8** | **48.9** | **30.5** | **49.3** | **62.3** |
| Swin-B [23] | 107 | 496 | 46.9 | 69.2 | 51.6 | 42.3 | 66.0 | 45.5 | 98 | 477 | 45.0 | 66.4 | 48.3 | 28.4 | 49.1 | 60.6 |
| Random-Swin-B | 107 | 496 | **49.1** | **71.6** | **64.2** | **44.6** | **68.2** | **47.4** | 98 | 477 | **47.4** | **68.8** | **50.7** | **31.3** | **51.6** | **64.2** |
| CSwin-T [8] | 42 | 279 | 46.7 | 68.6 | 51.3 | 42.2 | 65.6 | 45.4 | – | – | – | – | – | – | – | – |
| Random-CSwin-T | 42 | 279 | **47.3** | **69.3** | **51.6** | **42.8** | **66.4** | **45.9** | – | – | – | – | – | – | – | – |
| CSwin-S [8] | 54 | 342 | 47.9 | 70.1 | 52.6 | 43.2 | 67.1 | 46.2 | – | – | – | – | – | – | – | – |
| Random-CSwin-S | 54 | 342 | **48.8** | **71.3** | **53.4** | **44.0** | **67.9** | **47.3** | – | – | – | – | – | – | – | – |
| BiFormer-S [49] | 45 | – | 47.8 | 69.8 | 52.3 | 43.2 | 66.8 | 46.5 | 35 | – | 45.9 | 66.9 | 49.4 | 30.2 | 49.6 | 61.7 |
| Random-BiFormer-S | 45 | 265 | **48.4** | **70.6** | **52.4** | **43.7** | **67.2** | **46.9** | 35 | 246 | **46.5** | **67.4** | **50.0** | **30.4** | **50.6** | **62.3** |
| Pooling-Based Backbone | | | | | | | | | | | | | | | | |
| PVTv2-B1 [36] | 34 | – | 41.8 | 64.3 | 45.9 | 38.8 | 61.2 | 41.6 | 24 | – | 41.2 | 61.9 | 43.9 | 25.4 | 44.5 | 54.3 |
| Random-PVTv2-B1 | 33 | 216 | **43.0** | **66.1** | **46.9** | **39.1** | **52.4** | **42.9** | 23 | 197 | **42.4** | **63.0** | **44.8** | **26.6** | **45.2** | **57.1** |
| PVTv2-B2 [36] | 45 | – | 45.3 | 67.1 | 49.6 | 41.2 | 64.2 | 44.4 | 35 | – | 44.6 | 65.6 | 47.6 | 27.4 | 48.8 | 58.6 |
| Random-PVTv2-B2 | 41 | 252 | **47.1** | **68.4** | **51.0** | **42.4** | **65.8** | **45.6** | 31 | 233 | **46.0** | **66.6** | **49.2** | **28.6** | **50.0** | **59.9** |
| PVTv2-B3 [36] | 65 | – | 47.0 | 68.1 | 51.7 | 42.5 | 65.7 | 45.7 | 55 | – | 45.9 | 66.8 | 49.3 | 28.6 | 49.8 | 61.4 |
| Random-PVTv2-B3 | 56 | 342 | **47.9** | **69.3** | **53.0** | **43.5** | **66.2** | **47.3** | 46 | 323 | **47.0** | **68.1** | **49.9** | **29.8** | **50.6** | **62.5** |

# [CVPR2026] *Random Wins All: Rethinking Grouping Strategis for Vision Tokens.*

- **Why the random better?**
  - Position Information
  - Head Feature Diversity
  - Large Receptive Field
  - Fix Random Pattern

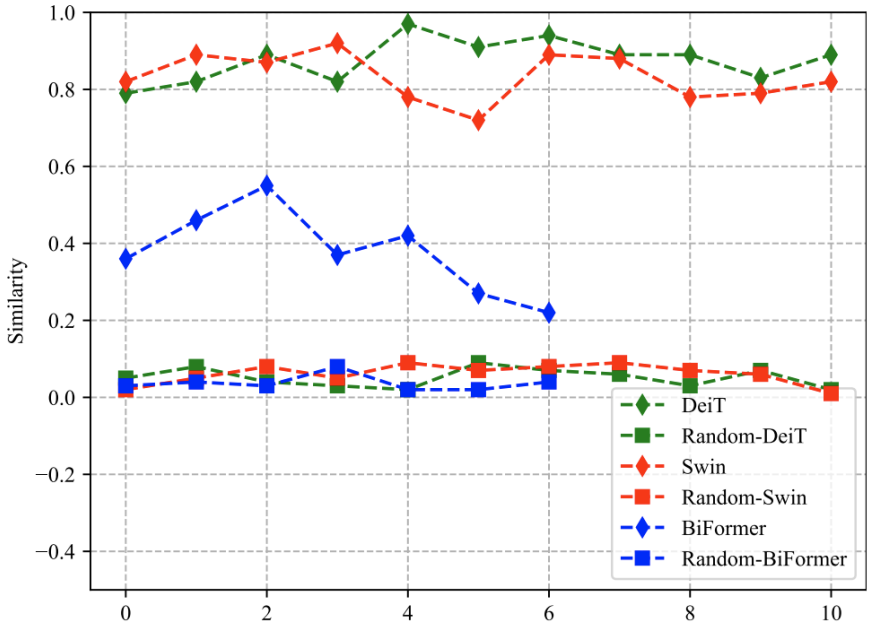
# [CVPR2026] *Random Wins All: Rethinking Grouping Strategis for Vision Tokens.*

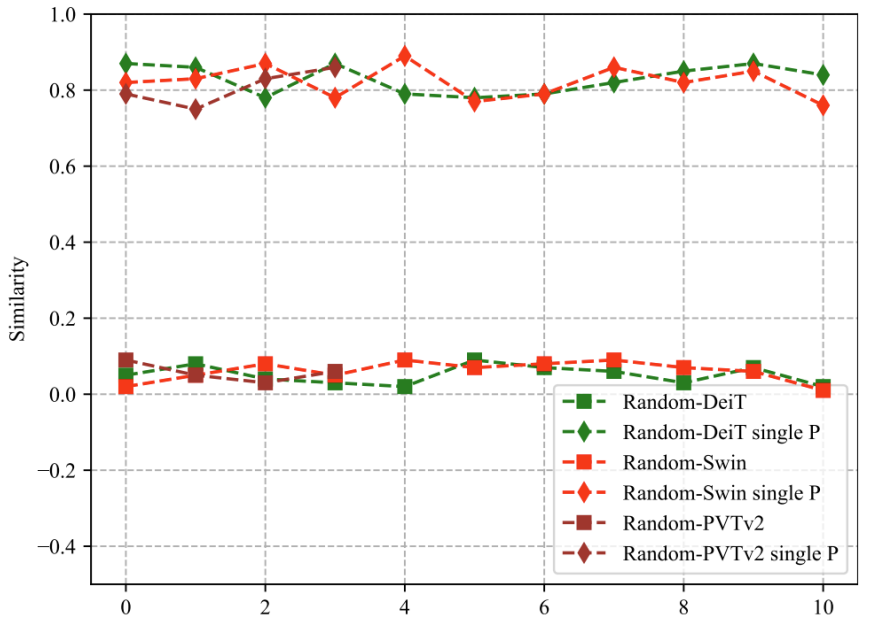
- **Why the random better?**
  - Position Information



# [CVPR2026] *Random Wins All: Rethinking Grouping Strategis for Vision Tokens.*

- **Why the random better?**
  - Head Feature Diversity
  - Metric: $\mathrm{Sim}(X_m, X_n) = \frac{\sum_{i=1}^{N} \cos(X_{mi}, X_{ni})}{N}$
  - Results:



# [CVPR2026] *Random Wins All: Rethinking Grouping Strategis for Vision Tokens.*

- **Why the random better?**
  - Large Receptive Field **&& Fixed Grouping Pattern**
  - Randomly shuffling the tokens gives the model an opportunity to perceive more distant tokens.
  - **The fixed random array caused the model to learn a fixed sparse pattern**.

| Model | Params(M) | FLOPs(G) | Acc(%) |
|---|---|---|---|
| DeiT-S [32] | 22 | 4.6 | 79.8 |
| Random-DeiT-S | 22 | 4.3 | 80.9 |
| global→region | 22 | 4.3 | 79.3(-1.6) |
| Swin-T [23] | 28 | 4.5 | 81.3 |
| Random-Swin-T | 28 | 4.5 | 82.7 |
| global→region | 28 | 4.5 | 81.5(-1.2) |
| PVTv2-B1 [36] | 13 | 2.1 | 78.7 |
| Random-PVTv2-B1 | 12 | 2.2 | 79.6 |
| global→region | 12 | 2.2 | 79.3(-0.3) |

| Model | Params(M) | FLOPs(G) | Acc(%) |
|---|---|---|---|
| DeiT-S [32] | 22 | 4.6 | 79.8 |
| Random-DeiT-S | 22 | 4.3 | 80.9 |
| Fully Random | 22 | 4.3 | 74.3(-6.6) |
| Swin-T [23] | 28 | 4.5 | 81.3 |
| Random-Swin-T [23] | 28 | 4.5 | 82.7 |
| Fully Random | 28 | 4.5 | 76.4(-6.3) |
| PVTv2-B1 [36] | 13 | 2.1 | 78.7 |
| Random-PVTv2-B1 | 12 | 2.2 | 79.6 |
| Fully Random | 12 | 2.2 | 75.4(-4.2) |

# Thanks!