Patch-Derived or Register Tokens? Comparing Variable-Length Image Tokenizers

Philipp Schenkel[1], Tobias Schürmann[1], and Ömer Şahin Taş[1,2]

[1]FZI Research Center for Information Technology [2]Karlsruhe Institute of Technology

Research Question

How does the choice between patch-derived and register tokens affect quality, efficiency, and spatial organization?

Flexible-length tokenizers represent images using **different token budgets**:

- **Patch-derived tokens** inherit spatial organization from image-patch features.
- **Register tokens** aggregate image information in learned embeddings, without a predefined spatial layout.

Existing approaches differ in architecture and training, making the **effect of the bottleneck design difficult to isolate**.

Our Approach

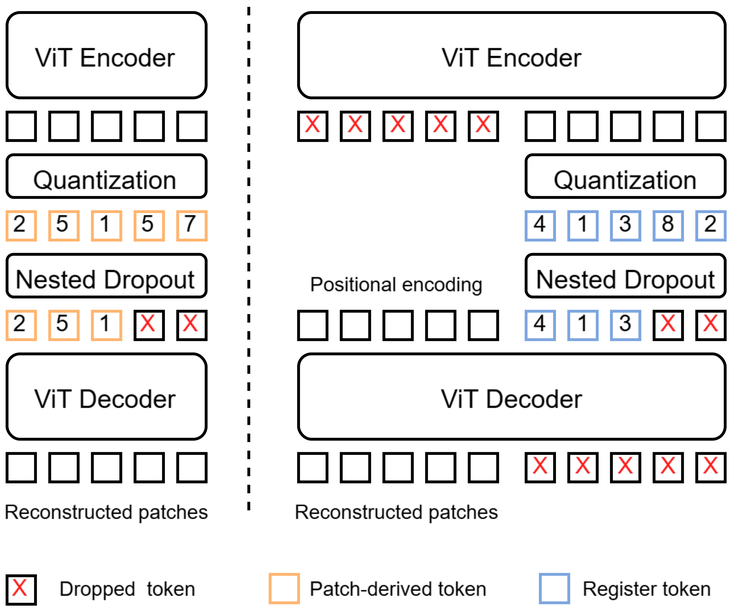
Controlled comparison: Same ViT backbone, quantizer, training objective, and training regime.

Variable length: Nested dropout retains token prefixes of varying lengths during training.

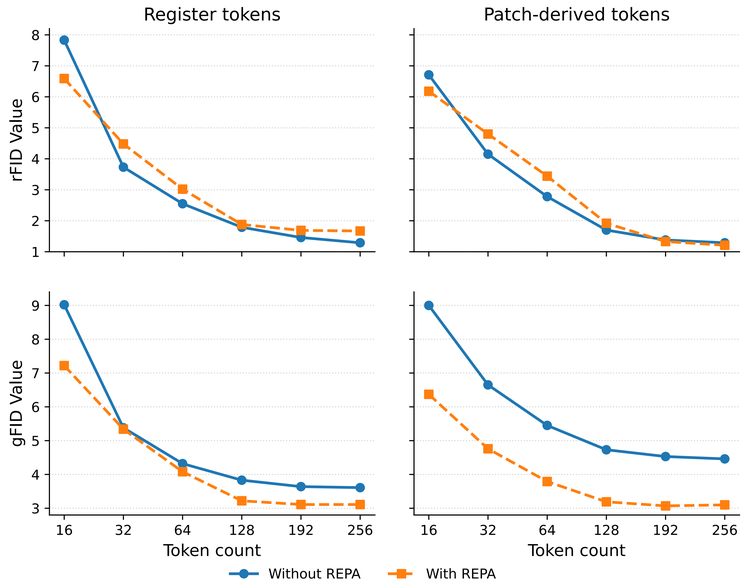
Representation alignment: Train both designs **with and without REPA**.

Evaluation: Compare **reconstruction quality, generation quality, and throughput** across token budgets; probe **spatial influence** by corrupting individual tokens.

Investigated Architectures



Reconstruction and Generation Metrics



Top: reconstruction FID. **Bottom:** generation FID. Lower is better.

Throughput Comparison

Token type	Encoder throughput	Decoder throughput				
		1	32	64	128	256
Patch-derived	1622.9	1756.8 (constant)				
Register	901.3	1750.9	1573.9	1452.6	1217.8	932.7

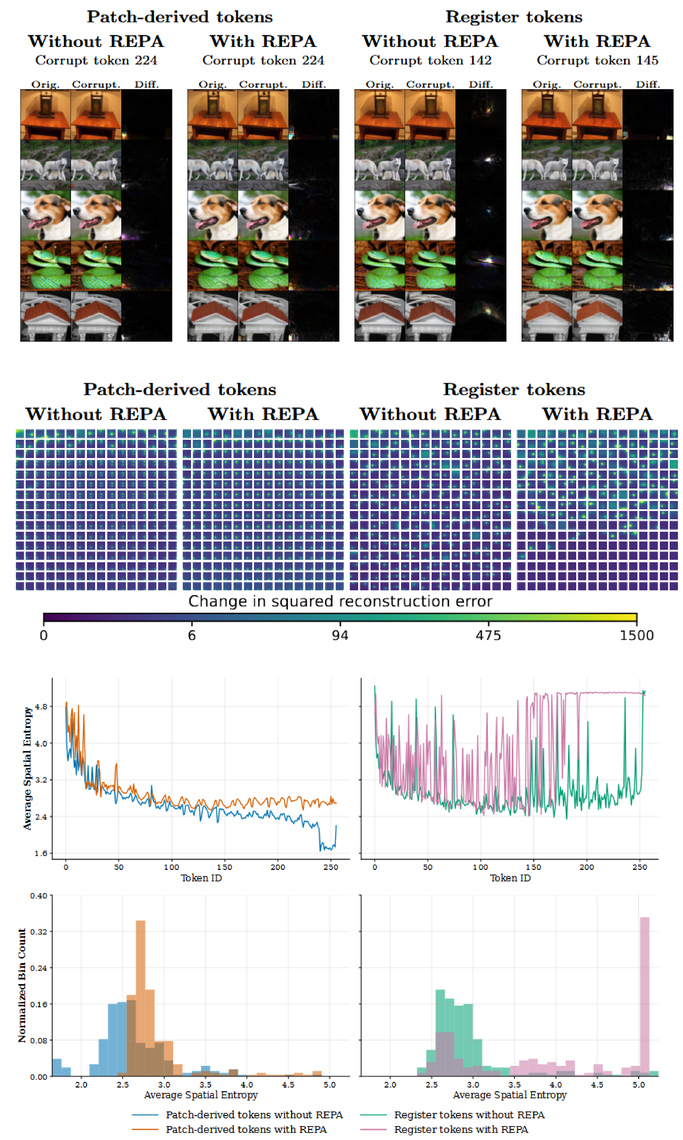
Images per second; decoder throughput is compared across token budgets.

Conclusion

- Both approaches enable **compact, variable-length** image representations.
- Patch-derived tokens achieve **higher throughput**, especially at high token counts.
- More tokens improve reconstruction and generation, with gains largely **saturating around 128–192 tokens**.
- **Without REPA**, register tokens achieve stronger downstream generative performance.
- **REPA narrows the generative-performance gap**, but its benefits depend on the downstream model's capacity.
- Patch-derived tokens exhibit **strong spatially localized influence**.
- Register tokens develop **spatial specialization** despite having no explicitly imposed two-dimensional layout.
- REPA **broadens spatial influence**, especially among register tokens.

Corruption Analysis

Replace one token at a time and measure the **change in the reconstruction**.



Open Questions

- Is the learned **spatial organization** desirable?
- What would a preferable **latent representation** look like?
- Can **higher compression rates** be achieved?