Exploring the Effectiveness of Object-Centric Representations in Visual Question Answering: Comparative Insights with Foundation Models

Amir Mohammad Karimi Mamaghan, Samuele Papa, Karl H. Johansson, Stefan Bauer, Andrea Dittadi

HELMHOLTZ MUNICH

Technische Universität München

Motivation

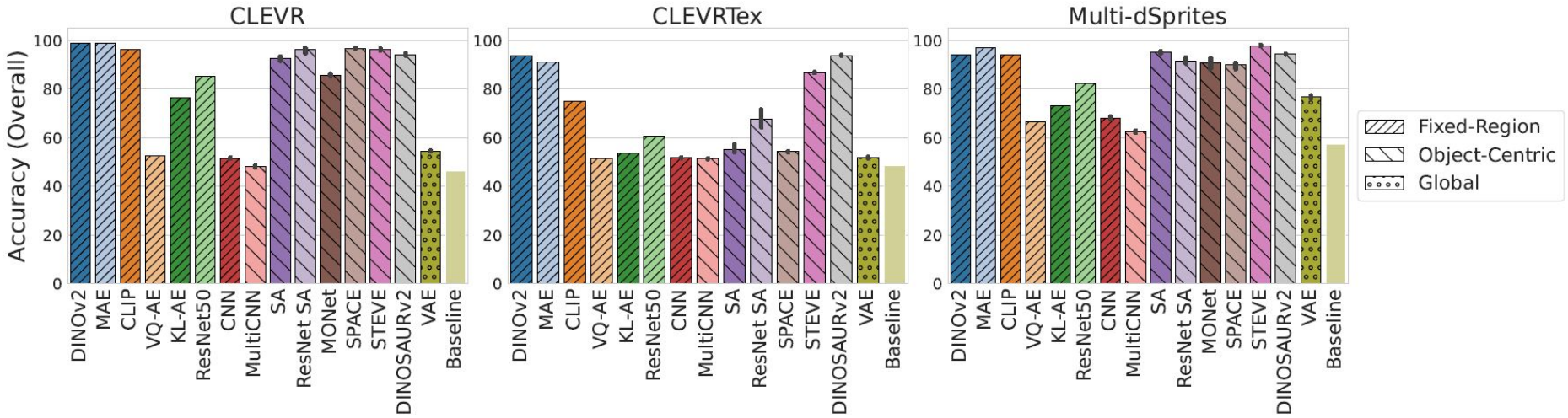
- Object-Centric Learning: A paradigm for visual representation learning that decomposes scenes into discrete objects
  - Goal: capture the compositional properties of the real world
- Common evaluation: Unsupervised Object Discovery (segmentation)
- OC representations inherently encode key properties of OC learning that go beyond segmentation

Seitzer, Maximilian, et al. "Bridging the Gap to Real-World Object-Centric Learning." The Eleventh International Conference on Learning Representations (2023).

Goal

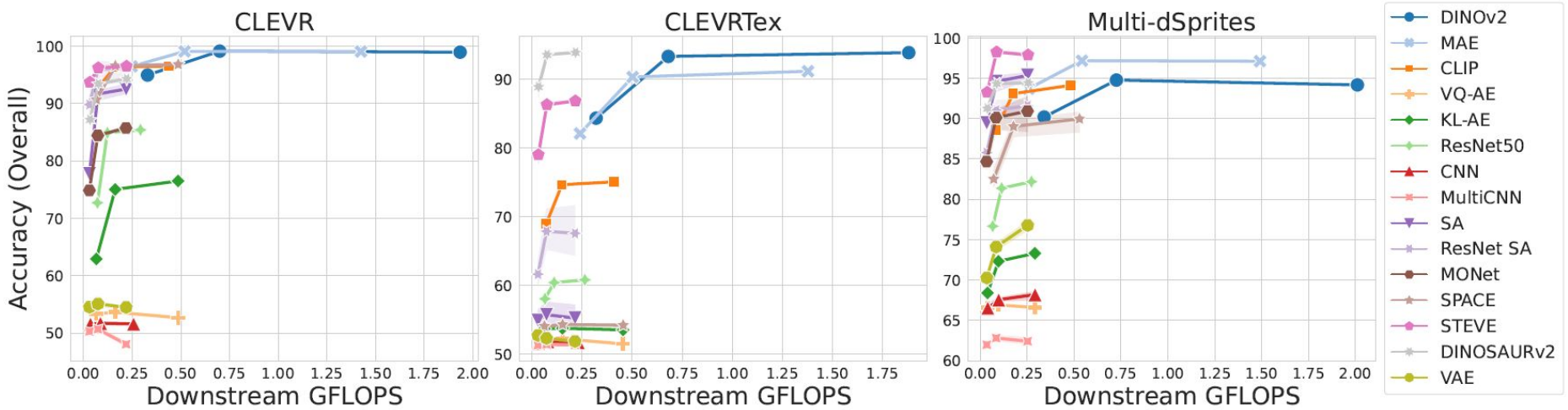
- In a visual reasoning (VQA) task at hand, which model should we use: an OC model or a foundation model? And what are the trade-offs of using each type of model?
  - Conduct a thorough empirical study on object-centric representations for downstream Visual Question Answering
  - Identify and investigate the trade-offs between large foundation models and object-centric models

VQA Performance



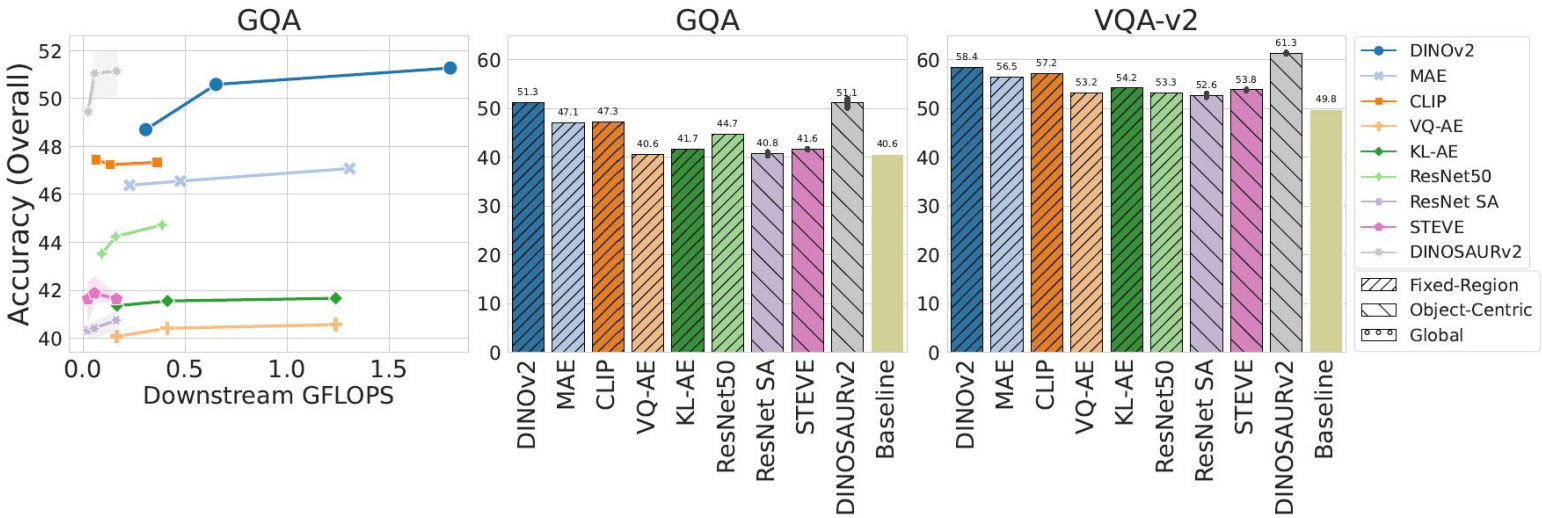
- FMs perform comparably to top-performing OC models without any training or hyperparameter tuning

Downstream Compute



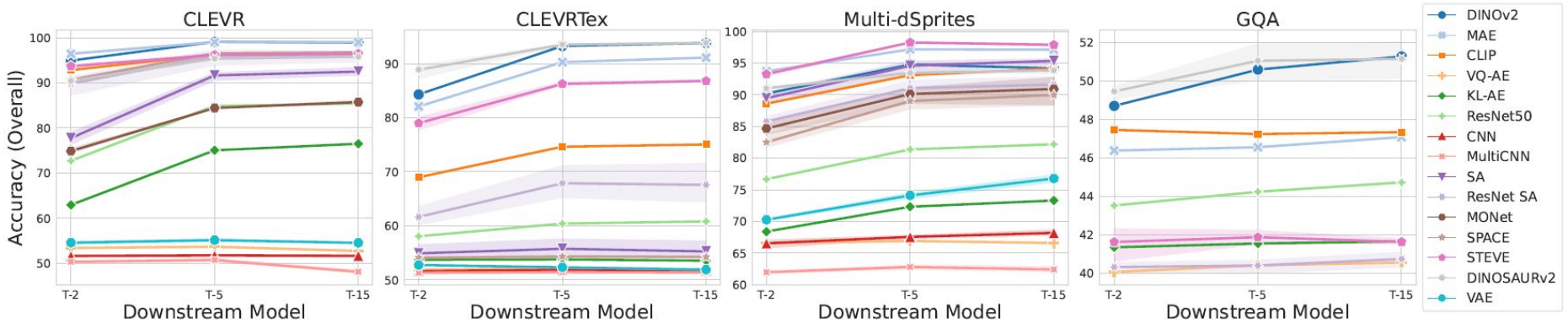
- FMs demand more compute
- FM + OC (DINOSAURv2) performs comparable to FM alone (DINOv2) while requiring much less downstream compute

Real-World Results



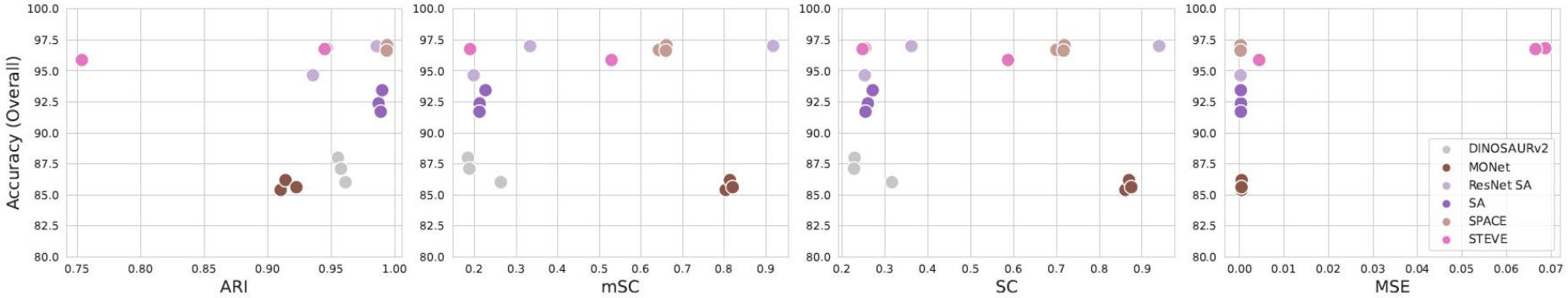
- Same patterns hold in real-world settings

Effect of Downstream Model Size



- DINOSAURv2 performs better than DINOv2 on smaller downstream models: More explicit representations

Upstream vs Downstream Performance



- Upstream metrics are not good predictors of downstream VQA performance
- Good segmentation ≠ Good representation

Conclusion

- Trade-offs of using a FM vs an OC model
  - FMs perform comparably off-the-shelf
  - FMs require more compute
  - FMs have less explicit representations
- FM + OC bias can be a viable solution to achieve the best of both worlds
  - Main downside: OC bias must be trained
- Importance of downstream evaluation
  - Most benefits of OC learning lie in the representations
  - Need both upstream (segmentation) and downstream (reasoning) evaluation