

A APPENDIX

A.1 COMPARISON WITH STATE-OF-THE-ART

Below, we list the benchmark results on ACRE (Table 3) and CATER (Table 4). Our proposed model is competitive with the state-of-the-art on CATER, as it underperforms compared to OPNet (Shamsian et al., 2020) by .7%. However, this difference can be justified by considering the task-engineered architectural design decisions behind OPNet. The OPNet model is composed of a perception module and two reasoning modules tailored specifically for the CATER task; one reasoning module determines what object to track, and another handles the target in the case of full occlusion and determines where its location is. This architecture would not translate well to other reasoning tasks, such as ACRE for example, where tracking and/or occlusions are not relevant. We can therefore, by design, expect OPNet to have a strong performance on CATER - and the fact that our model achieves competitive performance with the state-of-the-art while making no task-specific design decisions whatsoever is encouraging. Indeed, our model can extend to ACRE as well, with no architectural adjustment.

On the other hand, ALOE (Ding et al., 2021) is similar to our approach in that the reasoning module is modeled as a general transformer architecture that can be applied to multiple reasoning tasks without incorporating inductive biases. However, ALOE still relies on a separately trained perceptual model, and still resembles a two-stage pipeline. We improve on this by proposing a singular, unified architecture that learns strong implicit object-centric representations (as demonstrated by probing), while also being able to solve reasoning tasks. We find that we are able to match ALOE’s performance, empirically, while avoiding using any task-specific losses (like the L1 loss).

Our model is not hand-designed for a particular task, unlike OPNet, and does not require any separately trained perceptual models, unlike ALOE - given these improvements over the existing methods, we find it encouraging that we can still achieve strongly competitive performance in comparison.

Table 3: Benchmark results on ACRE. Numbers apart from our method are taken from the ALOE (Ding et al., 2021) paper. We show that our end-to-end unified architecture is able to achieve competitive results with the current state-of-the-art on causal reasoning.

Model	ACRE (Comp)
CNN-BERT	43.79%
NS-OPT	69.04%
Aloe	91.76%
Our Method	83.81%

Table 4: Benchmark results on CATER. Numbers apart from our method are taken from the ALOE (Ding et al., 2021) paper. We show that our end-to-end unified architecture is able to achieve competitive results with the current state-of-the-art on a complex spatiotemporal reasoning task.

Model	CATER Top 1 (Static)
R3D LSTM	60.2%
R3D + NL LSTM	46.2%
OPNet	74.8%
Hopper	73.2%
Aloe (no auxiliary)	60.5%
Aloe	70.6%
Aloe (with L1 loss)	74.0 ± 0.3%
Our Method	74.1%

A.2 PROBING

We further include bounding box and shape predictions for each of the ten slot embeddings in a frozen ResNet + Transformer encoder on 6 randomly sampled test frames from the LA-CATER dataset. These figures visualize what object-centric information each individual frozen slot embedding is encoding. As mentioned before, we notice each of the embeddings encode at most a few objects, and all of the objects in a scene are encoded by at least one embedding. Furthermore, we observe that the model is even able to recognize heavily occluded objects such as the small blue cube hidden behind the large blue cylinder in Figure 10, which is successfully captured by Slot Token 10.

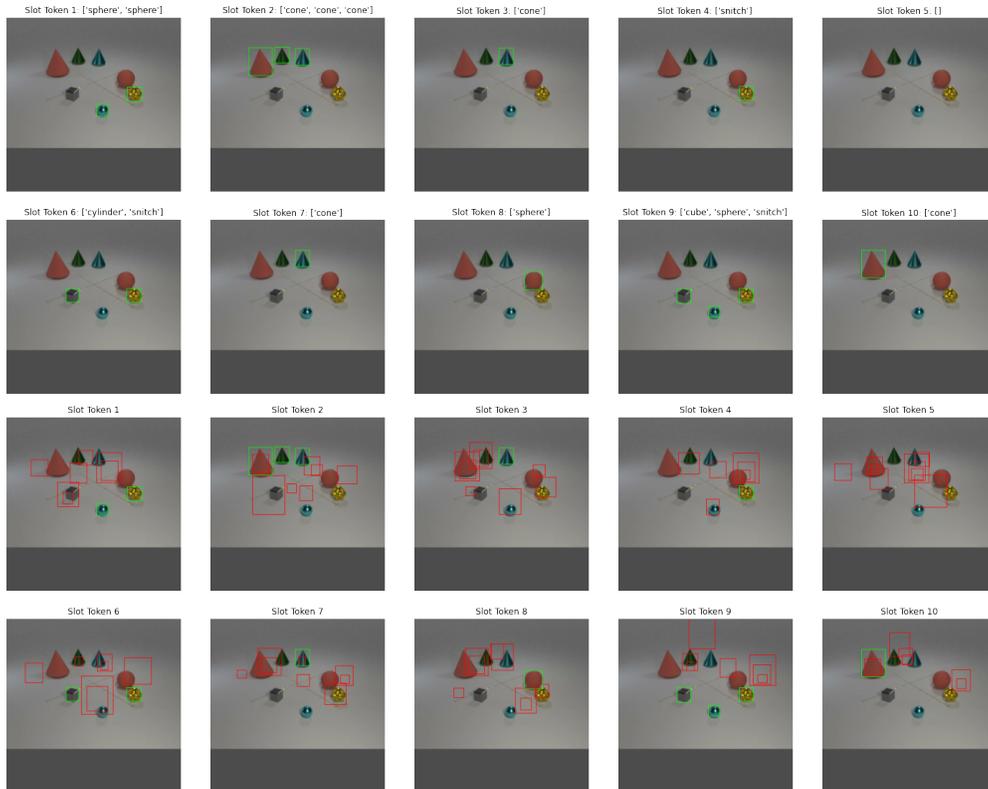


Figure 7: Sample test frame number 1 for probing from the LA-CATER dataset.



Figure 8: Sample test frame number 2 for probing from the LA-CATER dataset.

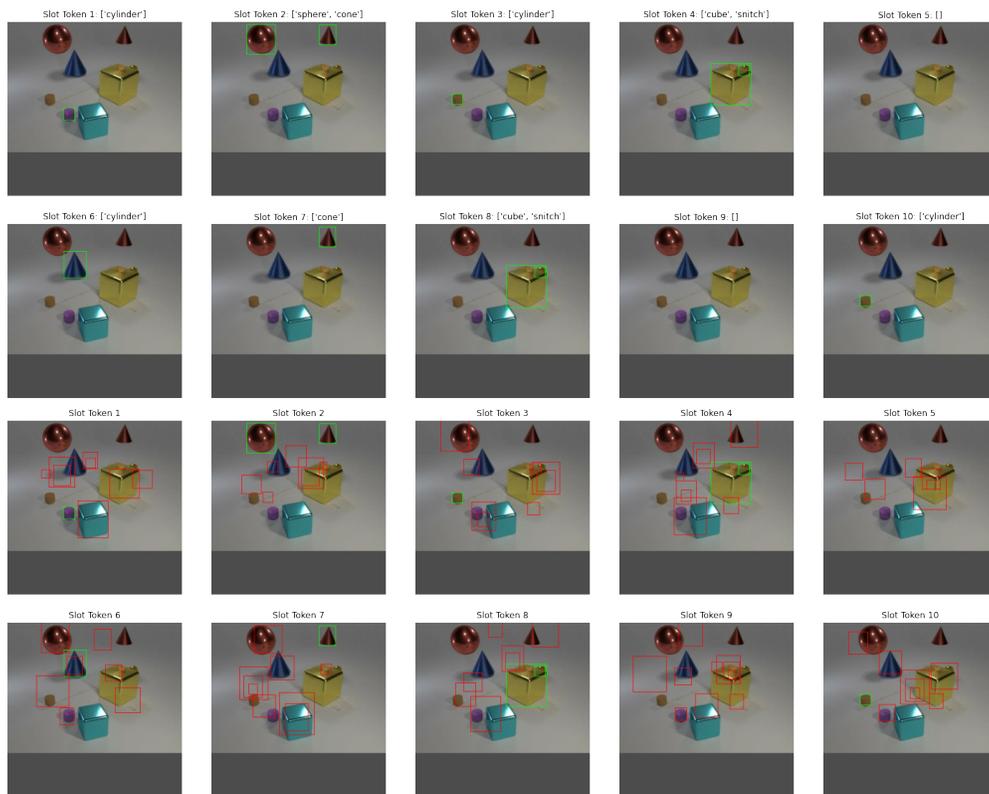


Figure 9: Sample test frame number 3 for probing from the LA-CATER dataset.

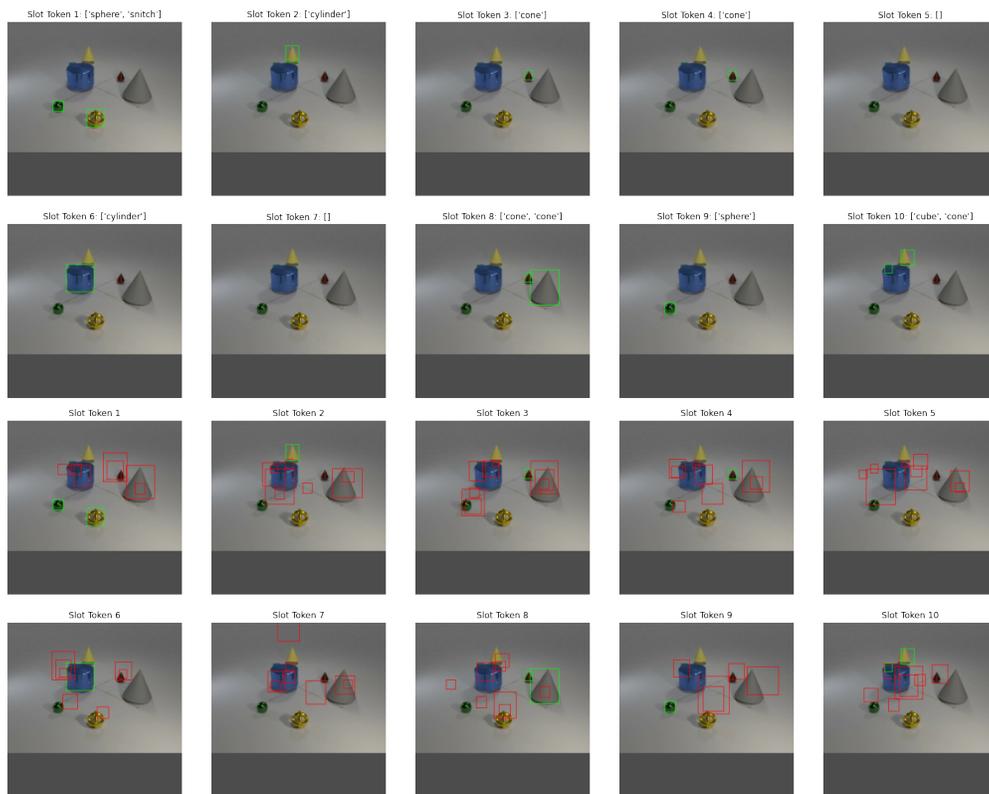


Figure 10: Sample test frame number 4 for probing from the LA-CATER dataset.



Figure 11: Sample test frame number 5 for probing from the LA-CATER dataset.

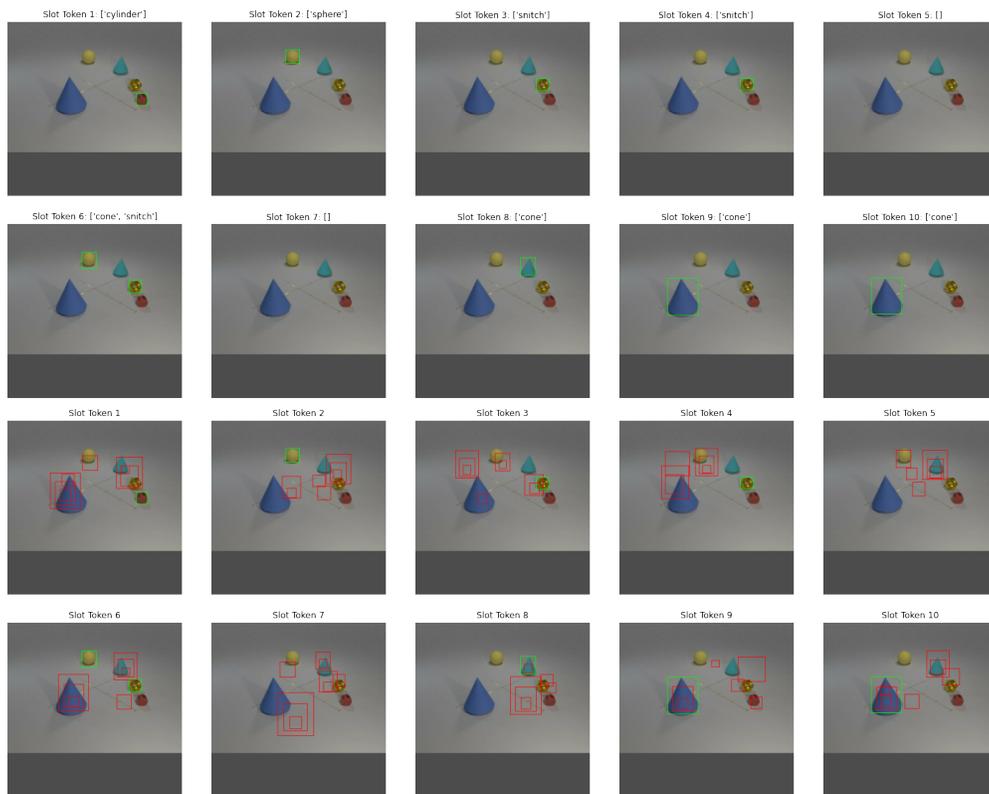


Figure 12: Sample test frame number 6 for probing from the LA-CATER dataset.