A Future Work

Fig. 9 and Fig. 10 both highlight intriguing qualitative differences between vanilla and implicit SLATE and understanding what causes these differences would be valuable for future work. What these figures highlight is that multiple decompositions of a scene into components are possible, which may differ in how closely they reflect human intuition on what constitutes a visual entity. This suggests that the optimization objectives for current object-centric models still underspecify the kinds of decompositions we seek to achieve in our models. The paradigm of decomposing static scenes, as opposed to interactive videos (e.g. [62, Fig. 7]), also contributes to this underspecification.

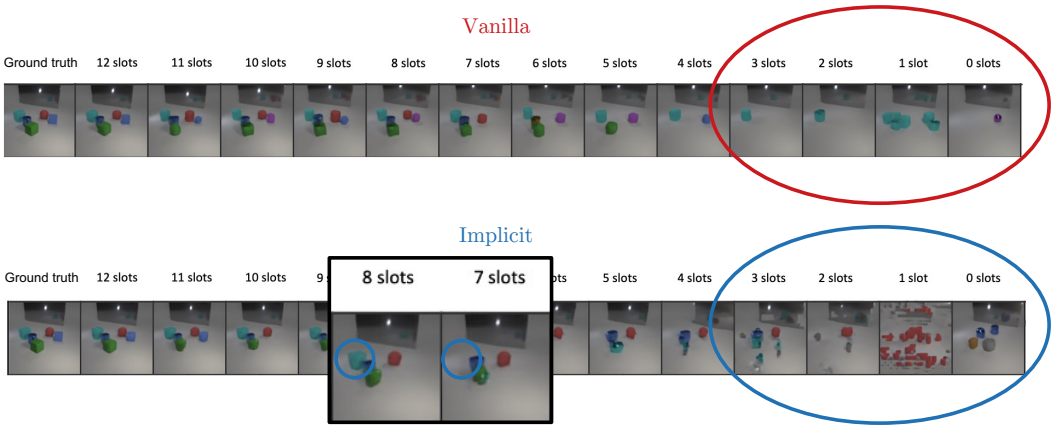
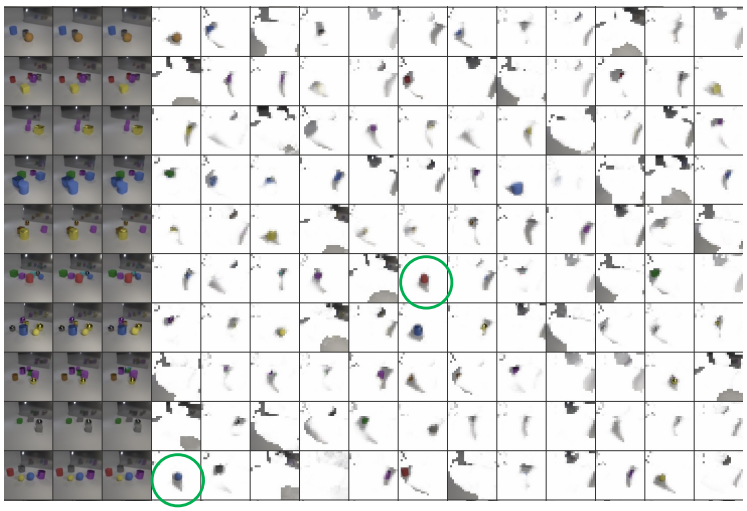


Figure 9: Implicit differentiation appears to create a stronger dependence among the slots. This figure shows what reconstruction looks if we train and evaluate with 12 slots, then re-render the reconstruction by deleting slots one at a time. When there are still many other slots as context, for both vanilla and implicit SLATE, deleting a slot corresponds to a clean deletion of the corresponding object in the reconstruction, as shown in the inset that highlights what the rendering looks like if we render with eight slots and seven slots. However, as we remove more slots, implicit SLATE generates less coherent compositions than vanilla SLATE, as shown when we render with only one to three slots. What causes this discrepancy is also an open question for future work.

Implicit SLATE



Vanilla SLATE

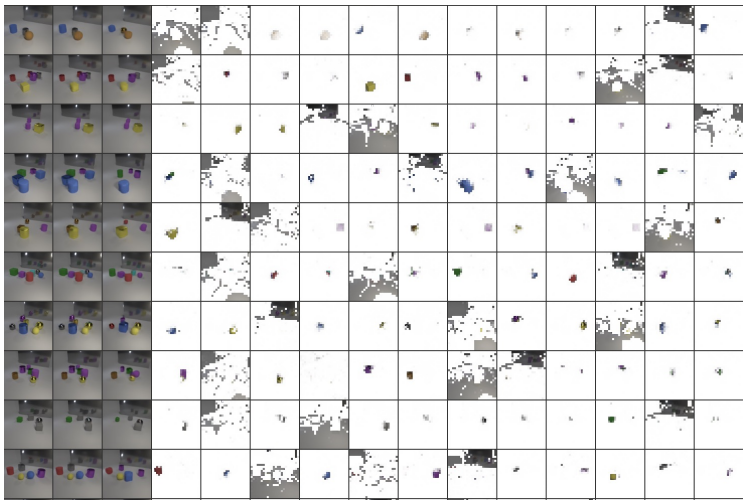


Figure 10: Despite our work pushing the optimization performance for a state-of-the-art model in object-centric learning (Tab. 3), and despite implicit slot attention producing similarly intuitive *predicted* segmentation masks as vanilla slot attention (Fig. 8), there appears to be a qualitative difference between the *attention* maps of implicit slate and those of vanilla slate. As this figure shows, the attention masks for vanilla SLATE appear to be more localized to each object, the attention masks for implicit SLATE appear to be more smeared out. One observation is that in some cases implicit SLATE appears to attend not only to the object but also its shadow, as circled in green. However, in other cases the attention maps appear to be smeared in other ways that may attend to a shadow that could possibly happen, but not necessarily a shadow in the given scene. What causes this discrepancy is open question for future work.

B Further Experiments

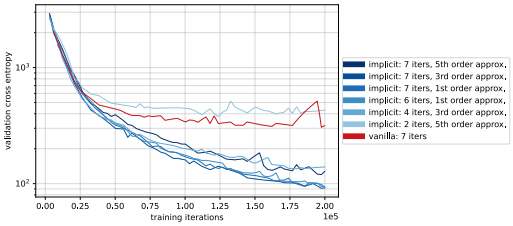


Figure 11: **Comparing different orders of Neumann approximation.** We sought to understand how the different orders of Neumann approximation affected performance. We observe that the 1st order approximation still largely performs the best, likely because adding more terms to the series expansion requires backpropagating through more iterations of slot attention, which was the problem we had sought to avoid in the first place. However, most approximations still perform better than the vanilla model with the same number of forward iterations.

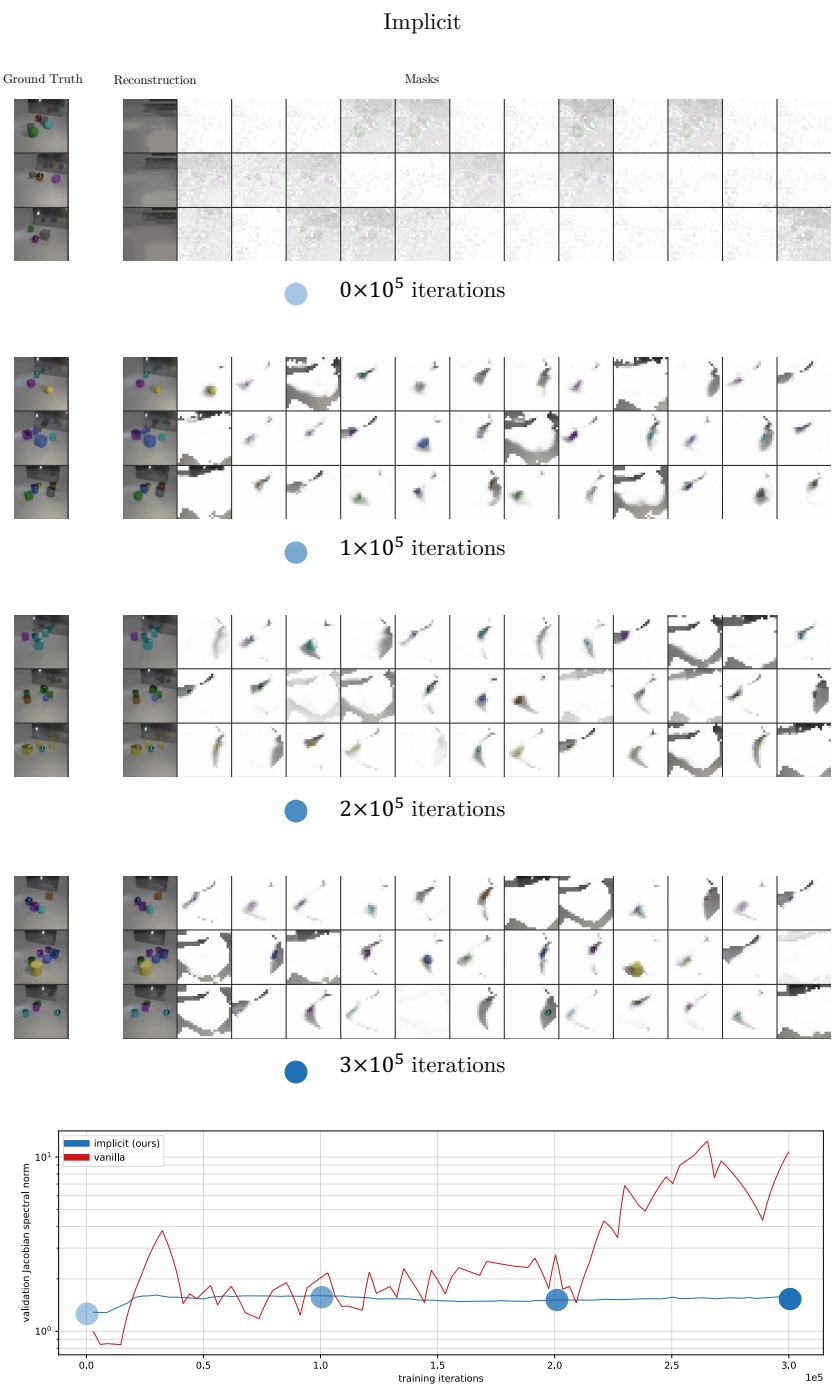


Figure 12: **Qualitative visualizations without gradient clipping: implicit.** This figure shows qualitative visualizations of implicit SLATE's reconstructions and attention masks when trained without gradient clipping. Compared to Fig. 13, implicit SLATE's reconstructions matches the ground truth much more closely, and its masks are more coherent, whereas vanilla SLATE's masks are much noisier, and become degenerate in the later stages of training as its Jacobian norm explodes.

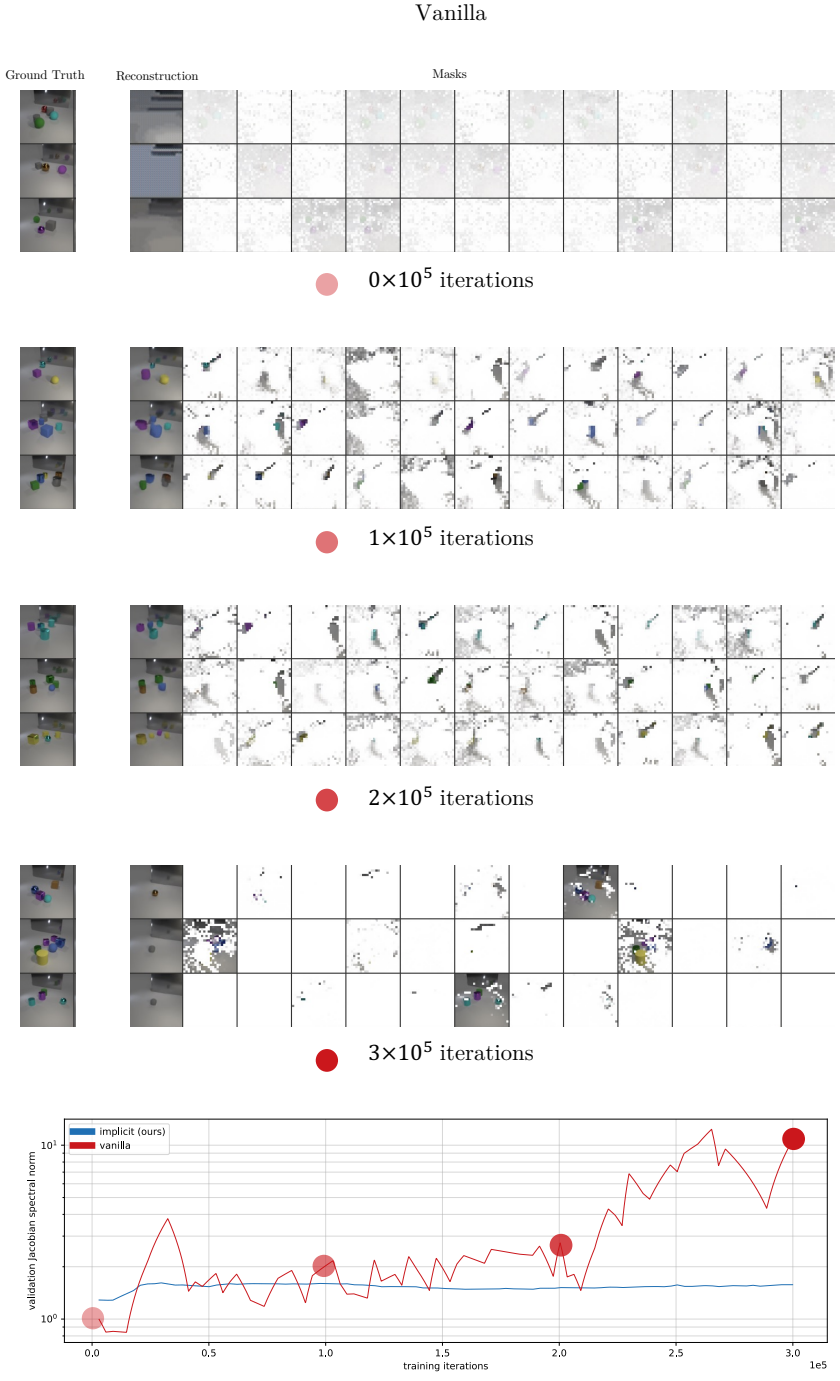


Figure 13: **Qualitative visualizations without gradient clipping: vanilla.** Compared to Fig. 12, vanilla SLATE's masks are much noisier, and become degenerate in the later stages of training as its Jacobian norm explodes, whereas implicit SLATE's reconstructions matches the ground truth much more closely, and its masks are more coherent.