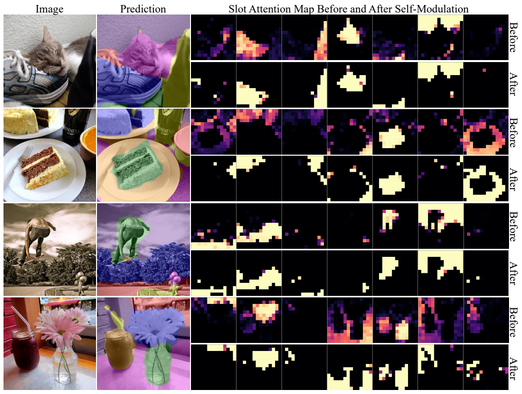

Figure R1: Visualization of attention maps of slot attention before and after self-modulation on COCO [27] (Fig. 3 in paper). The attention maps *before* self-modulation have been updated.

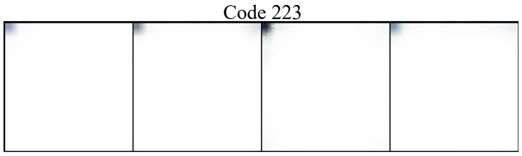

Figure R2: Visualization of a code trained on COCO [27] that associates with the top-left patch.

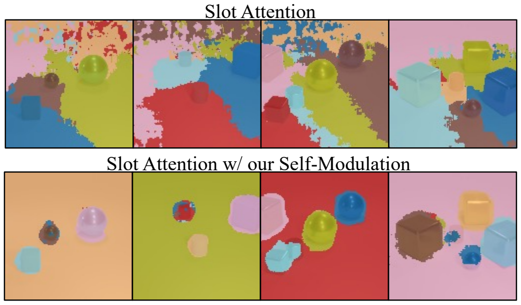

Figure R3: Visualization of predictions on CLEVR6 of slot attention [28] and slot attention with our self-modulation.

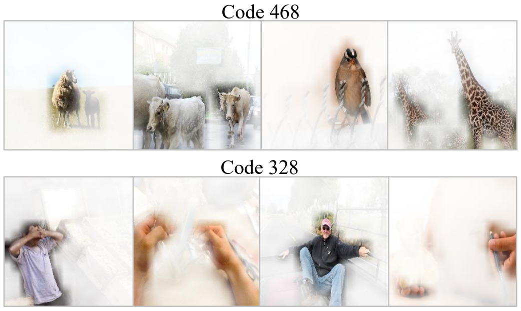

Figure R4: Visualization of the codebook on COCO [27]. The codebook also learns super-categories such as 'animals' and 'human'.

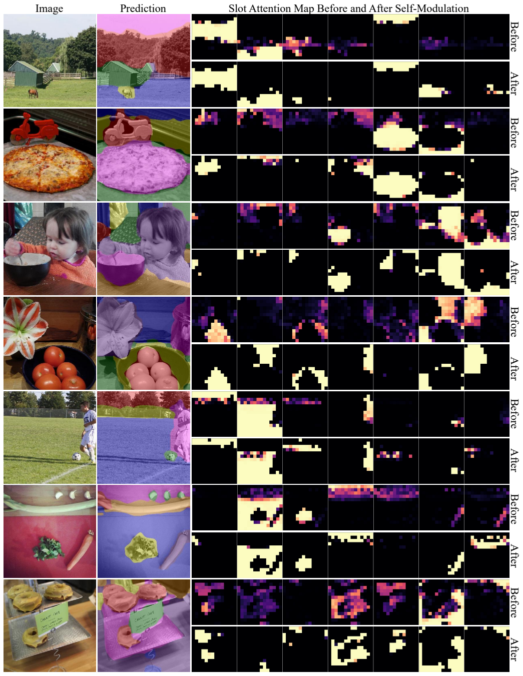

Figure R5: Visualization of attention maps of slot attention before and after self-modulation on COCO [27] (Fig. B.5 in paper). The attention maps *before* self-modulation have been updated.

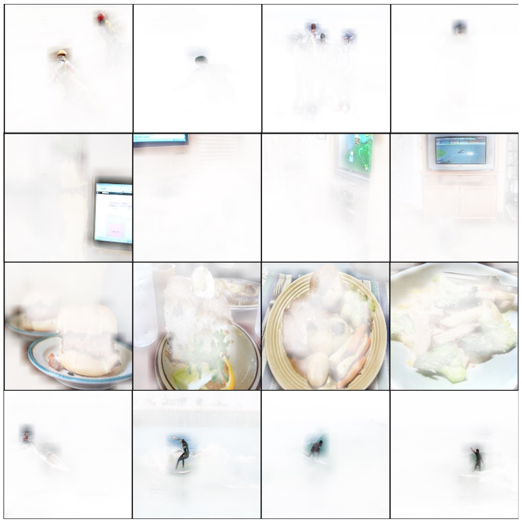

Figure R6: Visualization of K-means clusters on slot representations from DINOSAUR [30] on COCO [27]. The first cluster captures 'hat', the second 'electronic displays', the third 'plate holding food', and the fourth 'person enjoying water activities'.

Table R1: Comparison with DINOSAUR [30] on MOVI-C and -E [14] with corrected mIoUs.

Method	**MOVI-C**			**MOVI-E**		
	FG-ARI	mBO^i	mIoU	FG-ARI	mBO^i	mIoU
DINOSAUR [30]	55.7	42.4	-	-	-	-
DINOSAUR reprod	54.7±4.1	41.9±1.8	~~48.4±1.4~~ 41.0±2.1	53.8±2.1	34.5±1.7	~~36.7±1.7~~ 33.6±1.9
Ours	**58.9**±5.1	**46.8**±2.4	~~**53.1**±1.7~~ **45.9**±2.5	**59.7**±3.1	**39.3**±1.8	~~**41.3**±1.6~~ **38.3**±1.9