A Architecture Details

Tables 5, 6, 7, 8 describe hyperparameters for our experiments. In the case of using SMM instead of SA, we use an additional dimensionality reduction for slots via trainable matrix multiplication.

Table 5: Architecture of the CNN encoder for set property prediction and object discovery tasks experiments. The set prediction model uses a stride of 2 in the layers with *, while the object discovery model uses a stride of 1 in these layers.

Layer	Channels	Activation	Params
Conv2D 5 × 5	64	ReLU	stride: 1
Conv2D 5 × 5	64	ReLU	stride: 1/2*
Conv2D 5 × 5	64	ReLU	stride: 1/2*
Conv2D 5 × 5	64	ReLU	stride: 1
Position Embedding	-	-	absolute
Flatten	-	-	dims: w, h
LayerNorm	-	-	-
Linear	64	ReLU	-
Linear	64	-	-

Table 6: Spatial broadcast decoder for object discovery task on the CLEVR and ClevrTex datasets.

Layer	Channels/Size	Activation	Params
Spatial Broadcast	8 × 8	-	-
Position Embedding	-	-	absolute
ConvTranspose2D 5 × 5	64	ReLU	stride: 2
ConvTranspose2D 5 × 5	64	ReLU	stride: 2
ConvTranspose2D 5 × 5	64	ReLU	stride: 2
ConvTranspose2D 5 × 5	64	ReLU	stride: 2
ConvTranspose2D 5 × 5	64	ReLU	stride: 1
ConvTranspose2D 3 × 3	4	-	stride: 1
Split Channels	RGB (3), mask (1)	Softmax on masks (slots dim)	-
Combine components	-	-	-

Table 7: Spatial broadcast decoder for object discovery experiments on Tetrominoes and Multi-dSprites datasets.

Layer	Channels/Size	Activation	Params
Spatial Broadcast	64 × 64	-	-
Position Embedding	-	-	absolute
ConvTranspose2D 5 × 5	32	ReLU	stride: 1
ConvTranspose2D 5 × 5	32	ReLU	stride: 1
ConvTranspose2D 5 × 5	32	ReLU	stride: 1
ConvTranspose2D 3 × 3	4	-	stride: 1
Split Channels	RGB (3), mask (1)	Softmax on masks (slots dim)	-
Combine components	-	-	-

Table 8: Hyperparameters used for our experiments with the SLATE architecture.

Module	Parameter	Value
	Image Size	96
	Encoded Tokens	576
dVAE	Vocab size	4096
dVAE	Temp. Cooldown	1.0 to 0.1
dVAE	Temp. Cooldown Steps	30000
dVAE	LR (no warmup)	0.0003
Transformer	Layers	8
Transformer	Heads	8
Transformer	Hidden Dim.	192
SA/SMM	Num. slots	12
SA/SMM	Iterations	7
SA/SMM	Slot dim.	192

B Additional results for image reconstruction

Figure 6 shows validation cross-entropy curves during training.

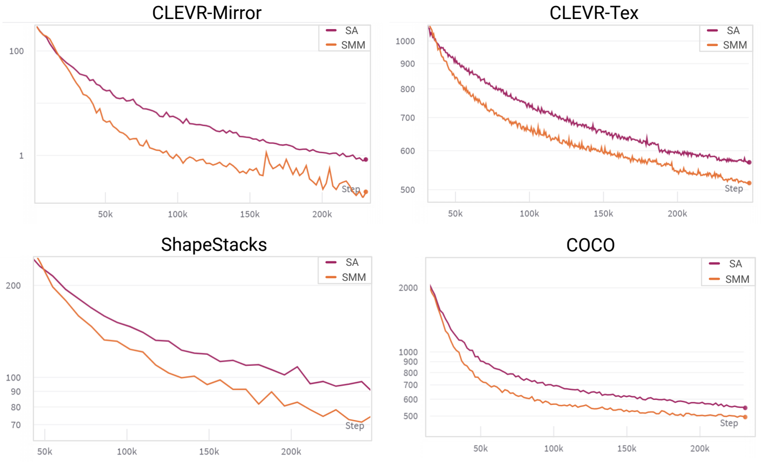

Figure 6: Validation cross-entropy during training for 4 different datasets. Our experiments show that using the SMM module instead of SA consistently improves the validation performance of the autoregressive transformer by about 10 percent during training. The result is maintained for all the datasets that we use.

C ADDITIONAL RESULTS FOR OBJECT DISCOVERY

Figure 7 shows examples of object-discovery on ClevrTex dataset for SMM and SA.

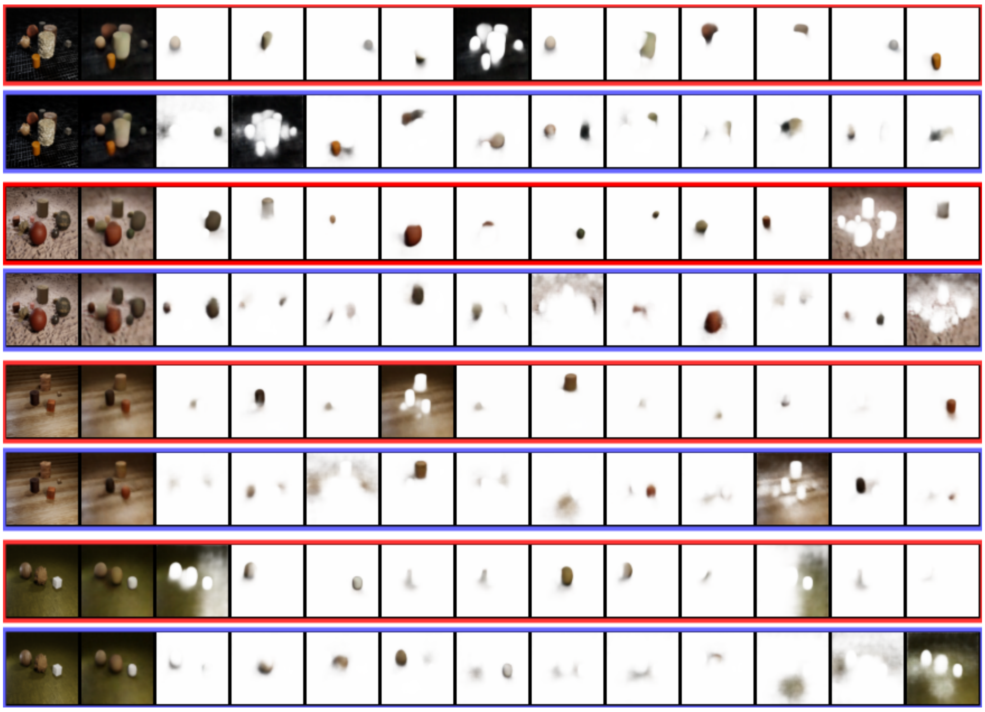

Figure 7: Examples of object-discovery on ClevrTex dataset for SMM and SA. The first column represents ground-truth images. The second one is broadcast-decoder reconstructions. The next columns are per slots of attention masks. Examples in the red borders are for SMM and examples in the blue borders are for SA.

D ABLATION STUDY

We also conducted the following ablation experiments. We trained Slot Attention (SA) and SMM object discovery models on the CLEVR6 dataset (scenes with six or fewer objects) with seven slots and three update iterations. Then we evaluated the trained models using different setups.

Evaluating the models on the CLEVR6 test split with 1, 2, . . . , and 7 update iterations gives the following FG-ARI scores (see Table 9).

Table 9: FG-ARI score for different number of update iterations (from 1 to 7)) at test time on the CLEVR6 dataset.

MODEL	1	2	3	4	5	6	7
SA	68.0	90.1	99.3	99.5	99.5	99.6	99.5
SMM	67.0	78.9	99.8	99.7	99.4	99.0	97.1

Evaluating the models on the CLEVR test images with only 7, 8, 9, and 10 objects (and a corresponding number of slots) gives the following results (see Table 10).

Table 10: FG-ARI score for different number of objects (from 7 to 10) at test time.

MODEL	7	8	9	10
SA	97.0	95.1	94.0	93.3
SMM	97.1	96.6	94.5	94.5

The results show that SMM is more robust to the number of out-of-distribution objects and less robust to different numbers of update iterations at test time.

E LIMITATIONS

The main limitations of SMM are due to the fact that the model belongs to the class of slot models, since it is based on a prominent representative of this class - Slot Attention. These limitations are: the need to specify the number of slots in advance, thereby setting the maximum possible number of objects in the image; over/under segmentation based on the specified number of slots; as a result, poor quality with a large number of slots; poor results and not aligned with human perception segmentation of real data.

While we have enhanced real-world object-centric image generation, the overall quality remains subpar and the attention maps scarcely resemble human object-focused vision. Future research should address the challenge of scaling and refining these visual models for complex real-world images. Also, as shown in the ablation study, SMM is less robust to different numbers of update iterations at test time.

We also want to draw attention to the potential negative societal impact of object-centric models, since modification and replacement of objects represented by a slot in an image can be used for malicious purposes.