Break-A-Scene

- □□□□□Break-A-Scene: Extracting Multiple Concepts from a Single Image
- □□□□□https://arxiv.org/abs/2305.16311
- SIGGRAPH Asia 2023

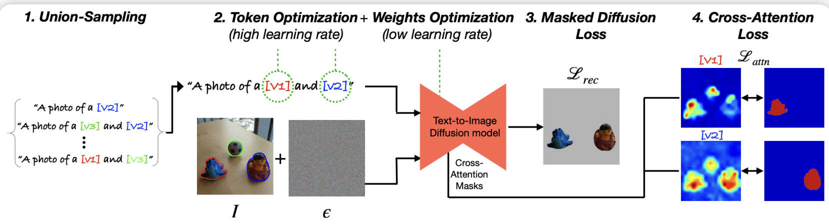

Figure 3: Method overview: our method consists of four key components: (1) in order to train the model to support different combinations of generated concepts, we employ a *union-sampling* mechanism, where a random subset of the tokens is sampled each time. In addition, (2) in order to avoid overfitting, we use a *two-phase training regime*, which starts by optimizing only the newly-added tokens, with a high learning rate, and in the second phase we also train the model weights, using a lower learning rate. A *masked diffusion loss* (3) is used to reconstruct the desired concepts. Finally, (4) in order to encourage disentanglement between the learned concepts, we use a novel *cross-attention loss*.

□□□token□□□mask□□□□□□□□□□mask□□embedding□□□□□□□□□□embedding□□□□□□

□□□□□mask diffusion loss□token□□□□□□□□□□□□□□□□□□□□cross-attention maps□□loss□□□□□□□□□□□□□□□□□□□□□□□□□□□union-sampling□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□N□□□□□□□□□□□

□□□□□□□□□□1□□□□□□□□□□□□□□□□□□□2□□□□□□□□□□□□□□□□3□□□4□□□□□□□□□□4□□□□□□□□□□

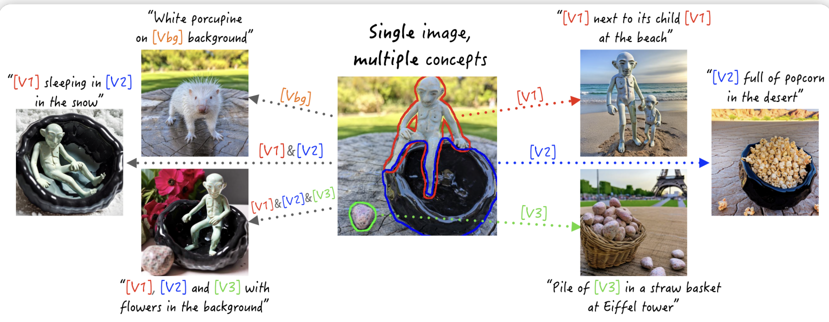

Figure 1: Break-A-Scene: Given a *single* image with *multiple* concepts, annotated by loose segmentation masks (middle), our method can learn a distinct token for each concept, and use natural language guidance to re-synthesize the individual concepts (right) or combinations of them (left) in various contexts.

- □□□test-tuning
- □□□prompt similarity□CLIP□□ID preservation DINO□□□DreamBooth□
- □□□□□□
- □□□https://github.com/google/break-a-scene