Supplementary Materials for

Shepherding Slots to Objects: Towards Stable and Robust Object-Centric Learning

Jinwoo Kim[1*] Janghyuk Choi[2*] Ho-Jin Choi[2] Seon Joo Kim[1]
[1]Yonsei University [2]KAIST
{jinwoo-kim, seonjookim}@yonsei.ac.kr
{janghyuk.choi, hojinc}@kaist.ac.kr

1. Limitations

We point out some limitations of the present work which could be possibly settled in the future works.

Large-scale encoder and decoder. In recent works [13–15] for OCL, Transformer [17] is adopted as an encoder [13] or decoder [14, 15] to handle the complex scenes. In step with these works, expanding SLASH with a Transformer-like encoder and decoder is a promising path for tiding over high-resolution images with complex real-world scenes.

Slot communication. As the SLASH directly inherits the Slot Attention [11], our model has a similar limitation with the Slot Attention, that is, the absence of communication between slots in the Slot Attention module. SLASH has another module which could be improved with slot communication: Point Predictor in IPPE. The Point Predictor in SLASH yields 2D coordinates without considering the relationship between slots. A communicable predictor where the model prevents redundant or omitted points can be devised in future works.

Background-related modeling. Our method does not care about any inductive bias or specific modeling for the background. As we found out that the background noise induces training instability in OCL, understanding background with external inductive bias can directly help models to be trained in a consistent way. We expect this to be done by a novel architecture with an additional criterion or ground truth signal to teach models the background knowledge.

Slot Initializer. Our model still contains the previous slot initializer using the Gaussian distribution with the learnable mean and sigma. We expect that replacing this slot initializer with a new one having some inductive biases concerning OCL can prevent noisy attention maps generated in the early iteration of the Slot Attention module.

	mIoU	ARI	fg-ARI
		CLEVR10	
SA [11]‡	36.6 ± 24.8	—	95.9 ± 2.4
MONet [1]‡	30.7 ± 14.9	—	54.5 ± 11.4
IODINE [5]‡	45.1 ± 17.9	—	93.8 ± 0.8
GenV2 [3]‡	9.5 ± 0.6	—	57.9 ± 20.4
SLASH*	49.7 ± 6.0	82.0 ± 14.2	94.9 ± 1.2

Table 1. Results of the object discovery task (mean ± std reported in %). ‡ stands for the scores recorded in [10] with 3 trials and data-specific center crop. The scores of SLASH is calculated with 10 trials.

2. Broader Impact

In this work, we adopt inductive biases and a weak semi-supervision scheme to achieve stable and robust OCL on the top of the Slot Attention [11]. Since the Slot Attention is a generic soft k-mean clustering algorithm, there is room for applying our approach to various domains where data has compositional characteristics. Not limited to vision tasks, it is expected that the motivation of our method can be applied to various domains and tasks, such as speech decomposition [12, 16], or music source separation [2, 6].

3. Experiments

3.1. Additional Dataset: CLEVR10

We provide additional results on the CLEVR10 dataset. As shown in Tab. 1, SLASH outperforms in all metrics, recording the highest average scores and low standard deviations, representing robustness and stability. Unlike the other models, our model solves the task without any data-specific pre-processing, i.e. center crop. Note that, in the CLEVR, as objects are crowded in the middle of images, this center crop can benefit the training and testing of the models.

	mIoU	ARI	fg-ARI
		CLEVRTEX	
SA ($\tau = 1$) [11]	22.2 ± 4.3	38.1 ± 12.5	52.1 ± 5.9
SA ($\tau = 2$)	25.6 ± 2.0	39.6 ± 3.7	54.9 ± 2.7
SA ($\tau = 5$)	19.0 ± 4.0	25.8 ± 11.6	48.3 ± 4.3
SA-LRE (0.25)	28.6 ± 9.8	44.2 ± 19.7	62.3 ± 12.6
SA-LRE (0.5)	27.2 ± 10.8	41.6 ± 22.5	59.7 ± 14.5
Gau ($\sigma = (0.01, 1.0)$)	25.2 ± 4.9	38.8 ± 12.9	58.9 ± 4.8
Gau ($\sigma = (0.1, 2.0)$)	26.0 ± 8.5	43.5 ± 14.6	55.9 ± 11.0
Gau ($\sigma = (1.0, 5.0)$)	24.7 ± 3.1	42.1 ± 7.4	55.2 ± 2.8
Conv	24.8 ± 6.0	42.5 ± 9.7	54.3 ± 11.1
WNConv (3×3)	28.8 ± 6.6	47.4 ± 14.3	58.0 ± 4.8
WNConv (5×5)	31.4 ± 6.6	55.6 ± 13.2	57.8 ± 7.7
WNConv (7×7)	29.0 ± 5.0	50.1 ± 15.4	59.6 ± 5.6
		PTR	
SA ($\tau = 1$)	17.6 ± 14.7	19.6 ± 29.8	44.5 ± 18.8
SA ($\tau = 2$)	34.3 ± 12.0	56.6 ± 26.5	50.0 ± 7.9
SA ($\tau = 5$)	33.7 ± 16.4	47.3 ± 32.4	56.7 ± 10.1
SA-LRE (0.25)	32.4 ± 19.1	40.1 ± 35.1	61.5 ± 7.0
SA-LRE (0.5)	35.5 ± 15.6	54.5 ± 34.1	62.1 ± 4.5
Gau ($\sigma = (0.01, 1.0)$)	22.6 ± 13.1	25.7 ± 8.4	56.4 ± 6.4
Gau ($\sigma = (0.1, 2.0)$)	20.6 ± 15.1	20.0 ± 29.9	53.8 ± 10.2
Gau ($\sigma = (1.0, 5.0)$)	20.2 ± 14.7	28.3 ± 31.4	58.4 ± 3.9
Conv	12.4 ± 9.7	11.6 ± 13.1	32.1 ± 26.0
WNConv (3×3)	41.4 ± 11.1	60.5 ± 21.2	60.3 ± 3.5
WNConv (5×5)	43.8 ± 3.0	62.3 ± 19.4	60.4 ± 3.2
WNConv (7×7)	40.2 ± 5.0	58.7 ± 24.3	61.0 ± 3.6

Table 2. Results of object discovery on the possible alternatives for ARK (mean ± std for 10 trials, reported in %). τ is the temperature in an attention mechanism. SA-LRE stands for Slot Attention with Low Resolution Encoder with the ratio of downsampling. Gaussian model (Gau) is given with the standard deviation σ with its min and max value. We use $\sigma = (0.1, 2.0)$ in the main paper as the default one. We mark WNConv (5×5), which is the setting for our method, as bold text.

3.2. Additional Alternatives for ARK

Tab. 2 describes an additional ablation study on the alternatives for the ARK. For the global smoothing technique, we include one more Slot Attention model with a temperature value of 5.0. In addition, the Slot Attention models with a low-resolution encoder, which we term SA-LRE, are added to the line of global smoothing methods.

For the alternative kernels of the WNConv, we provide the results of Gaussian smoothing with distinct standard deviations. We also conduct experiments on the different kernel sizes of WNConv, where we use 5×5 as a default one.

It can be easily observed that our proposed method and choice of the default WNConv show the most consistent and robust performance over the other alternatives.

3.3. Level of Semi-supervision

We conduct ablation studies on the level of semi-supervision according to two aspects: the number of images and objects. To inspect the sole impact of the weak semi-supervision, we only add the IPPE module to the Slot Attention model, i.e. SLASH without the ARK module. As shown in Tab. 3, we trained models by assigning ground truth annotations to 5, 10, 20, and 100% of images from a given dataset, and 50, 75, and 100% of objects in a given image. The results show that as the ratio of weak supervision ground truths grows, SLASH can perform better in the object discovery task. Furthermore, the full supervision, denoted as 100% and 100%, helps SLASH achieve the most consistent and high scores in both mIoU and ARI metrics.

imgs	objs	mIoU	ARI	fg-ARI
			CLEVRTEX	
5%		22.4 ± 3.1	34.2 ± 10.0	53.2 ± 6.8
10%	75%	25.1 ± 7.4	40.4 ± 15.6	54.9 ± 7.3
20%		26.7 ± 6.5	43.0 ± 7.1	56.9 ± 9.2
	50%	24.6 ± 5.2	41.3 ± 11.0	56.6 ± 7.8
10%	**75%**	25.1 ± 7.4	40.4 ± 15.6	54.9 ± 7.3
	100%	25.6 ± 8.9	41.5 ± 17.3	53.6 ± 7.3
100%	100%	27.5 ± 5.0	47.4 ± 10.5	57.7 ± 8.0
			PTR	
5%		32.6 ± 14.1	54.3 ± 33.0	51.7 ± 5.3
10%	75%	38.4 ± 12.8	58.4 ± 31.3	58.5 ± 3.1
20%		39.8 ± 14.2	59.4 ± 32.8	59.8 ± 3.2
	50%	34.9 ± 15.6	53.3 ± 32.7	58.2 ± 3.4
10%	**75%**	38.4 ± 12.8	58.4 ± 31.3	58.5 ± 3.1
	100%	38.6 ± 11.9	62.1 ± 27.6	58.1 ± 2.4
100%	100%	39.4 ± 10.8	66.5 ± 21.7	60.4 ± 8.2

Table 3. Results of ablation studies on the level of the semi supervision for the model of Slot Attention + IPPE which is shown as '+IPPE' in Tab. 2 in the main paper (mean ± std for 10 trials, reported in %). The ratio of the number of images and objects is described in the table. We mark the 10% for images and 75% for objects, which is the setting for our method, as bold text.

4. Qualitative Results

4.1. Intermediate and Final Results of SLASH

To investigate the overall process of SLASH, we visualize the intermediate attention maps of each slot in addition to the final results of the SLASH in Fig. 1, Fig. 2, Fig. 3, Fig. 4 for CLEVR6, CLEVRTEX, PTR, and MOVi, respectively.

Given an input (leftmost in the first row), SLASH iteratively updates the slots during which the attention maps between the slots and the feature vector are generated (first, third, and fifth rows). The attention maps before ARK depict our observation of salt-and-pepper patterns. Then ARK is applied to the attention maps to make the refined attention maps (second, fourth, and sixth rows). The noisy attention maps are cleansed by ARK so that the background noise is erased and the object pattern is strengthened. After $T = 3$ iterations, the decoder produces the segmentation masks (seventh row) and reconstruction images (last row).

4.2. Model Comparisons

In this section, we qualitatively compare the baseline Slot Attention (SA), a newly introduced weakly semi-supervised baseline WS-SA, and our model SLASH. For the fair comparison, we select the model which performs the second-best and the second-worst in the mIoU metric. The results of second-best and second-worst models for each dataset – CLEVR6 (Fig. 5, Fig. 6), CLEVRTEX(Fig. 7, Fig. 8), PTR(Fig. 9, Fig. 10), MOVi(Fig. 11, Fig. 12) – are illustrated below.

For the figure details, the first row contains the attention maps of K slots; $K = 11$ for CLEVRTEX, $K = 7$ otherwise. The second row contains the ground truth segmentation mask (leftmost), the aggregation of the predicted segmentation mask (second from the left), and the predicted segmentation masks for K slots. The last row contains the input image (leftmost), the final reconstructed image (second from the left), and the reconstructed images of K slots.

The results of the second-best models show that all the models perform well, where SLASH is superior to the other models in terms of robustness against background noise. The main difference comes from the results of the second-worst models. As mentioned in Sec. 4.4 in the main table, various types of the bleeding issue occur in the different datasets: irregular bleeding in CLEVR6, bleeding to background patterns in CLEVRTEX, the striping issue in PTR. In addition, for the MOVi dataset, we observe that the models, except for the SLASH, split the image into several blocks with less relation to objects. In contrast, even with the second-worst models, the SLASH accomplishes stable and robust outcomes across all datasets.

5. Implementation Details

5.1. Model

As our method is built on Slot Attention [11], the implementation of the Slot Attention module is just the same as [11]. Thus, in this section, we only describe the details for Attention Refining Kernel (ARK) and Intermediate Point Predictor and Encoder (IPPE).

ARK is a 5×5 single-channel single-layer convolutional kernel having only 25 learnable parameters and no bias. By applying `SAME` padding, the output size will be the same size as the input size.

IPPE has two submodules, Point Predictor and Point Encoder, consisting of 3-layer MLP with ReLU activation (Tab. 4). Point Predictor takes a slot of D_slot dimension and yields 2D coordinates (x, y). Point Encoder encodes a 2D point (x, y) into a vector of D_{slot} dimension. In our implementation, $D_{slot} = 64$ and $-0.5 < x, y < 0.5$.

Type	Size (Input/Output)	Activation
Point Predictor		
MLP	64/32	ReLU
MLP	32/16	ReLU
MLP	16/2	–
Point Encoder		
MLP	2/16	ReLU
MLP	16/32	ReLU
MLP	32/64	–

Table 4. Model details of IPPE.

5.2. Dataset

We use CLEVR dataset [8], also called CLEVR10, given by Multi-Object Datasets [9]. We split the CLEVR10 dataset into 70K and 15K for training and test set, respectively. We filtered out scenes containing more than six objects to compose CLEVR6 from the CLEVR dataset. As a result, in the CLEVR6 dataset, the training and test set have 35,050 and 7,492 images, respectively.

PTR dataset [7] is sourced from the official project page [1]. Also, for PTR, we utilize the validation set as a test set to use the ground truth segmentation masks for evaluation. CLEVRTEX dataset [10] is also sourced from the official project page [2]. We use part 1-4 for the training set and part 5 for the test set; we do not use the other variants of CLEVRTEX for our work.

For the MOVi dataset, we utilize the dataset generator provided by Kubric [4] in the official repository [3]. Although Kubric provides diverse variants of MOVi, we only focus on MOVi-C to -E for the following reasons: 1) MOVi-A and -B contains relatively simple objects compared to the other variants; 2) MOVi-F is for training optical flow predictor. Moreover, the only difference among MOVi-C, -D, and -E is in the dynamics of objects or camera view. We collected static images from MOVi-C, -D, and -E, and termed it MOVi or MOVi-C as there is no meaning to distinguish MOVi-C, -D, and -E in consideration of static image. In details, we set the number of static objects from 3 to 6 in a scene. For the clear observation of objects, we set the height of the camera greater than or equal to 5.0 in the half sphere whose radius ranges from 7 to 9. For data collection, only the first frames of the rendered videos are selected to prevent the existence of the same scenes in the dataset.

[1] http://ptr.csail.mit.edu
[2] https://www.robots.ox.ac.uk/ vgg/data/clevrtex/
[3] https://github.com/google-research/kubric

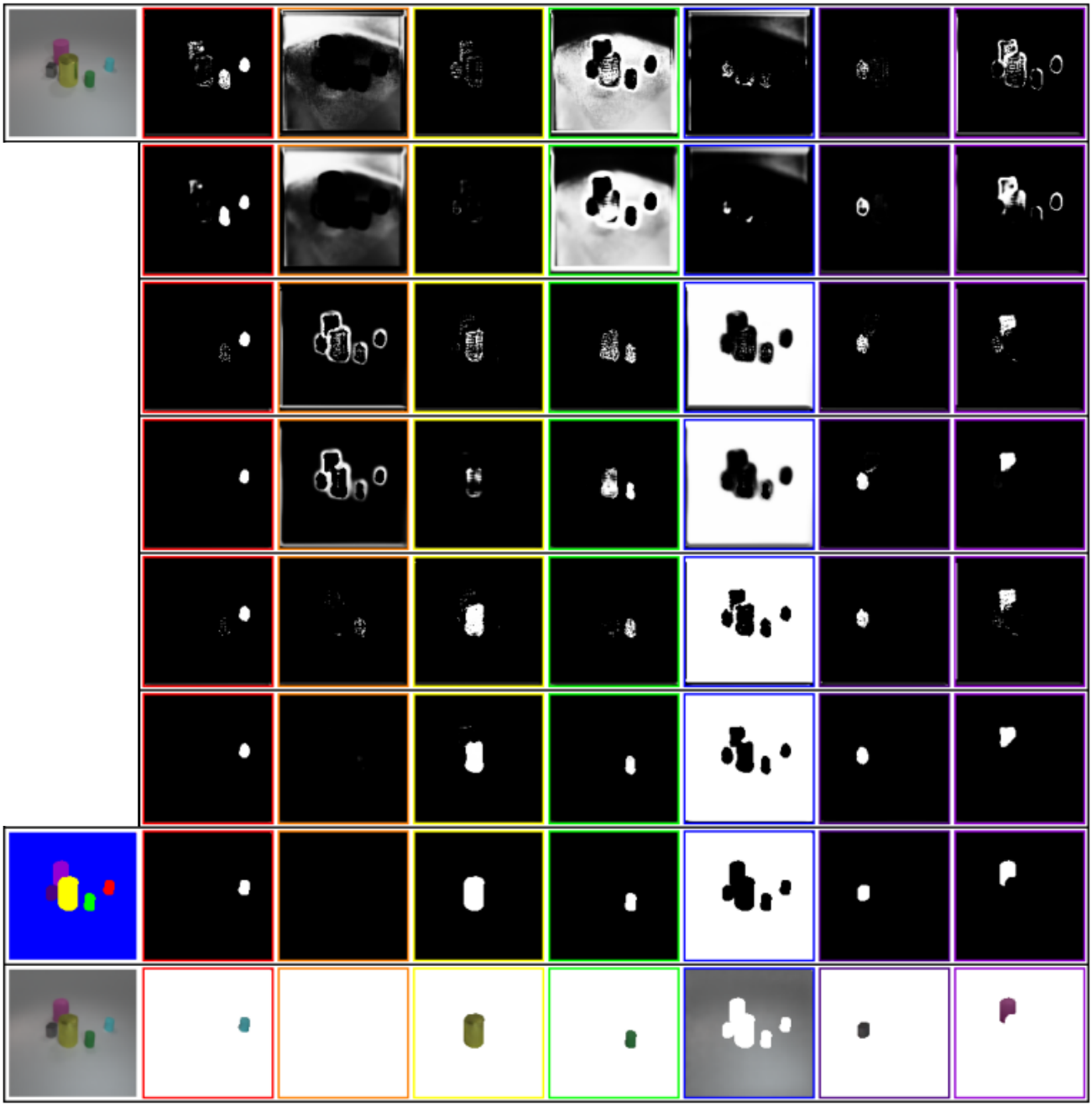

Figure 1. The intermediate and the final results of SLASH on CLEVR6. You can find the figure details in Sec. 4.1.

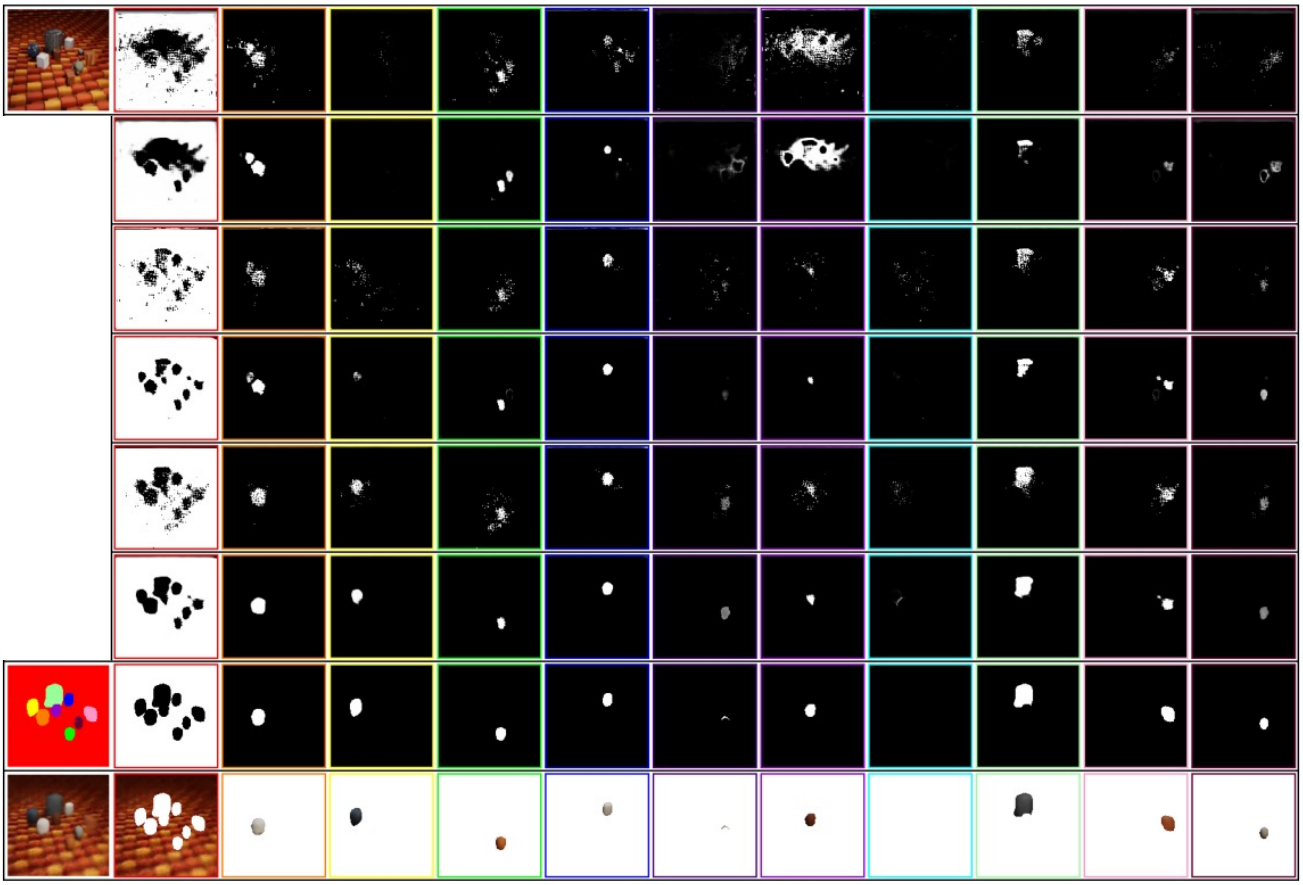

Figure 2. The intermediate and the final results of SLASH on CLEVRTEX. You can find the figure details in Sec. 4.1.

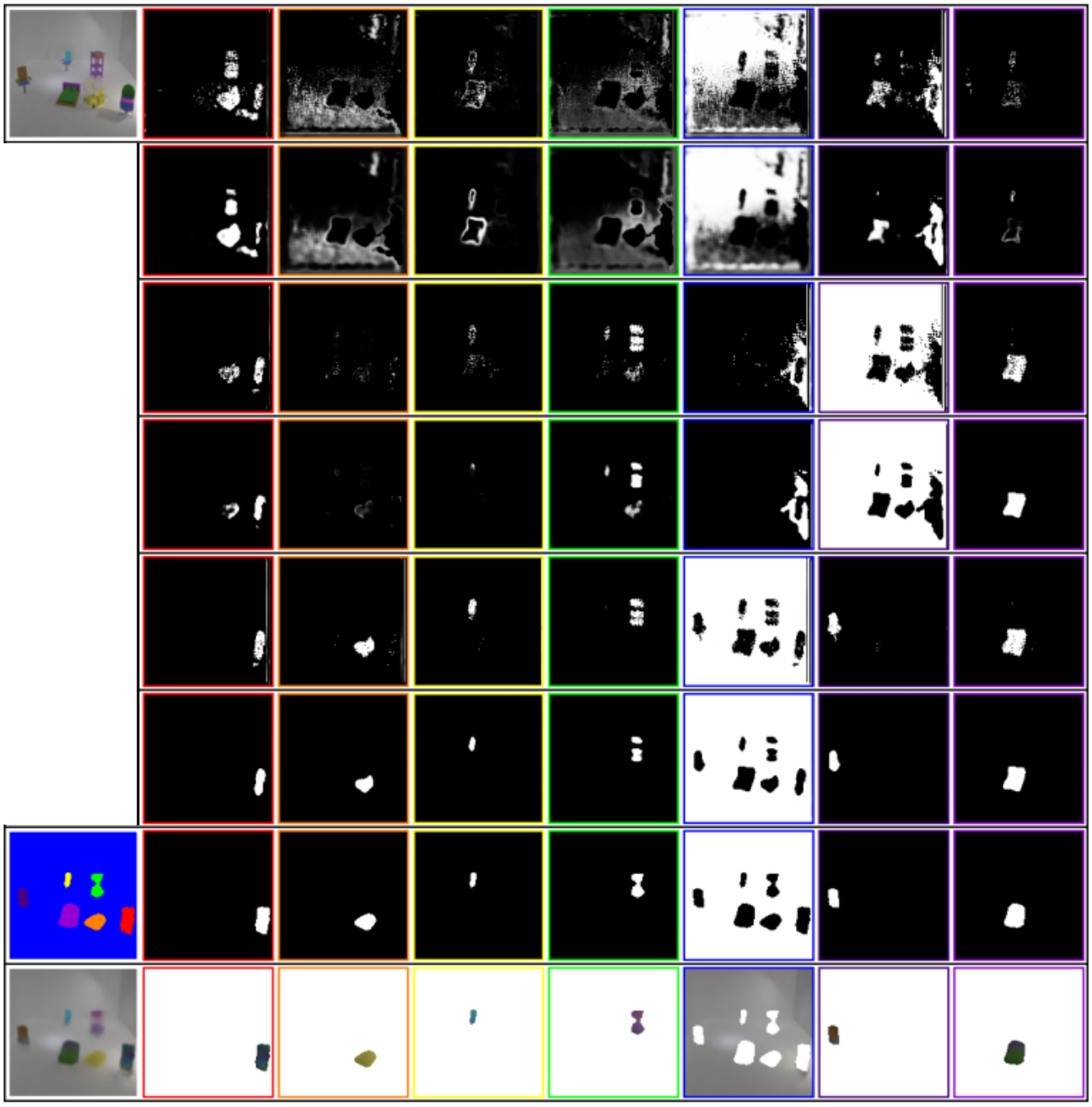

Figure 3. The intermediate and the final results of SLASH on PTR. You can find the figure details in Sec. 4.1.

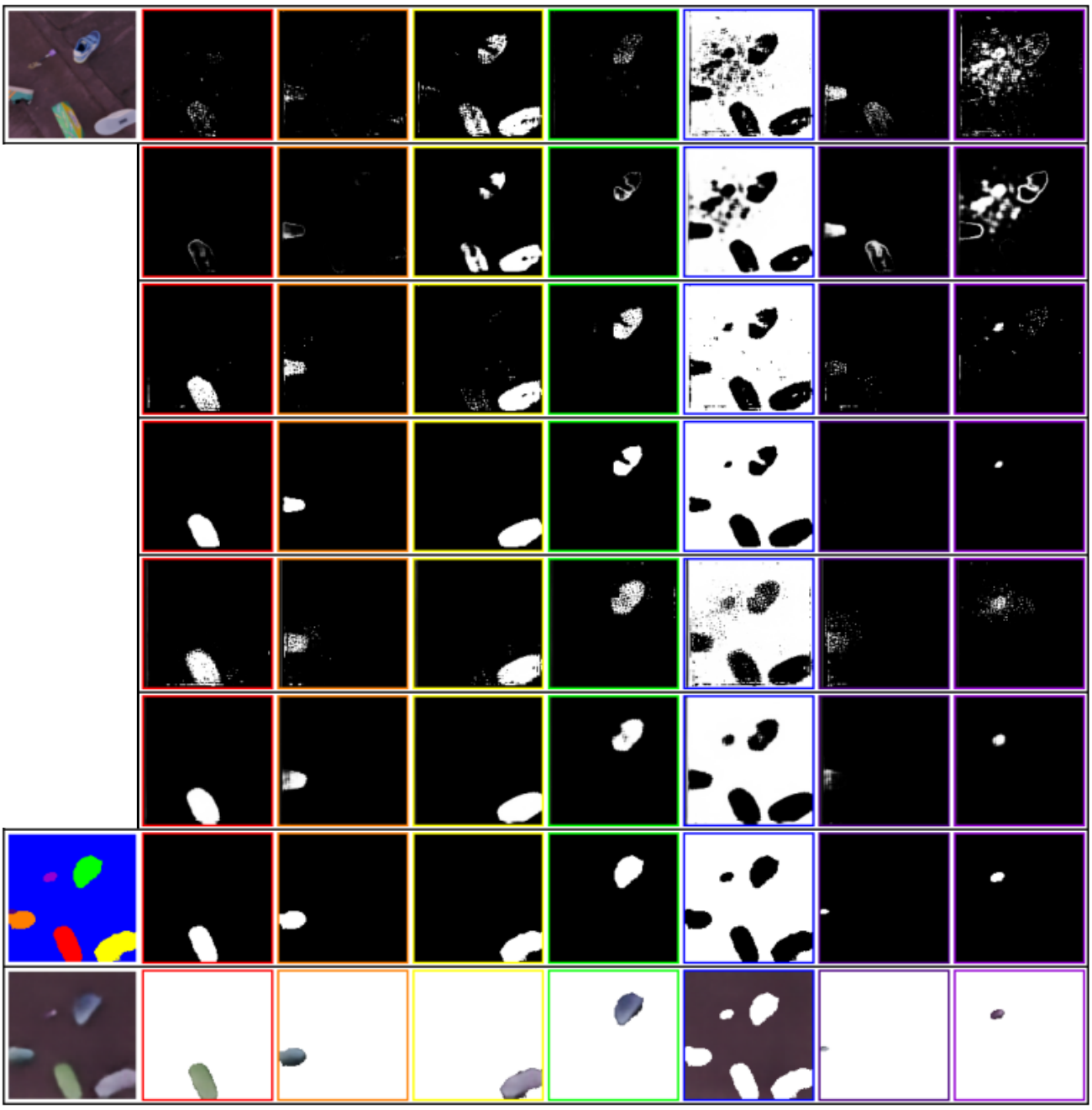

Figure 4. The intermediate and the final results of SLASH on MOVi. You can find the figure details in Sec. 4.1.

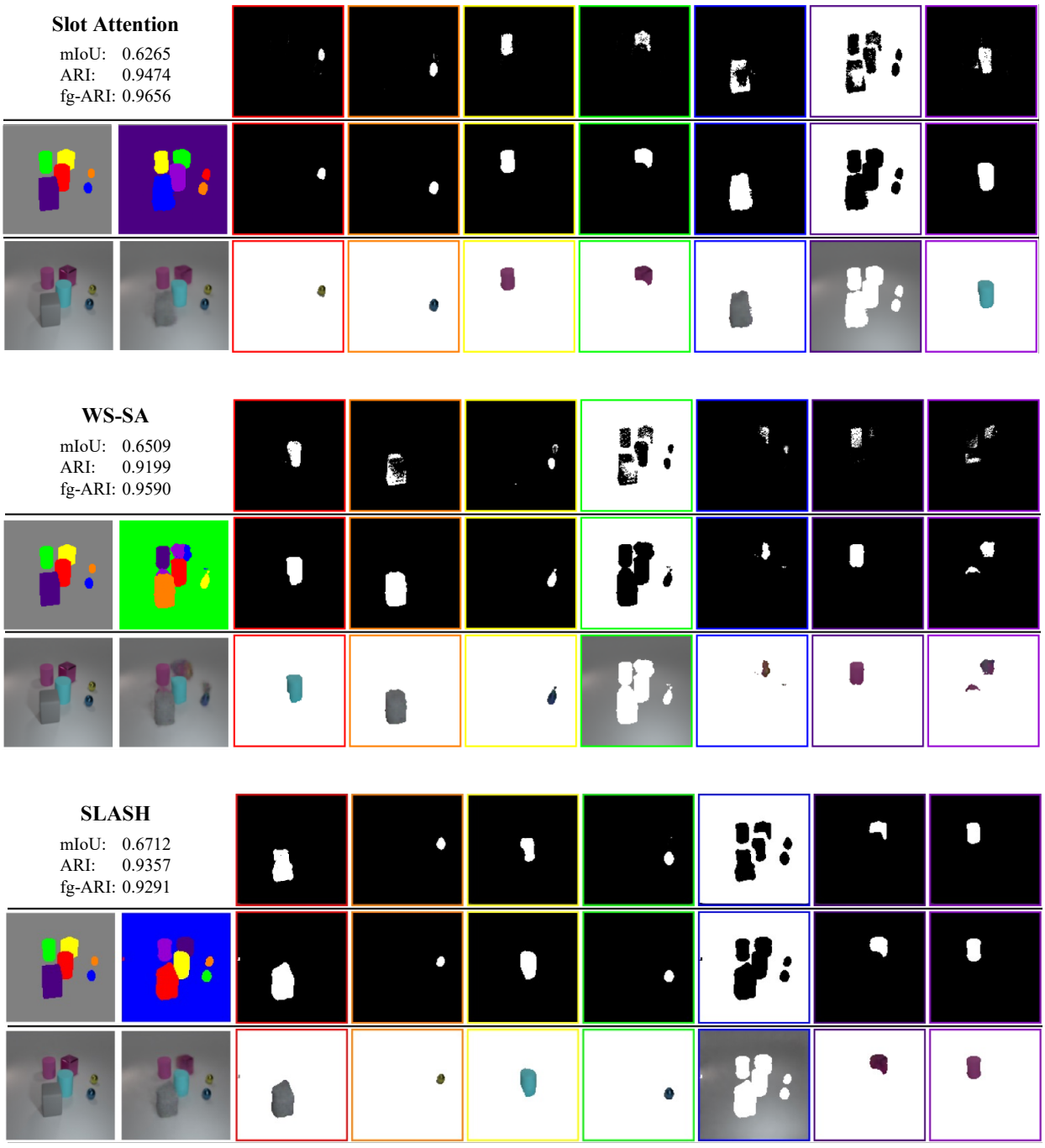

Figure 5. Results on CLEVR6 by the second-best models from SA, WS-SA and SLASH. You can find the figure details in Sec. 4.2.]

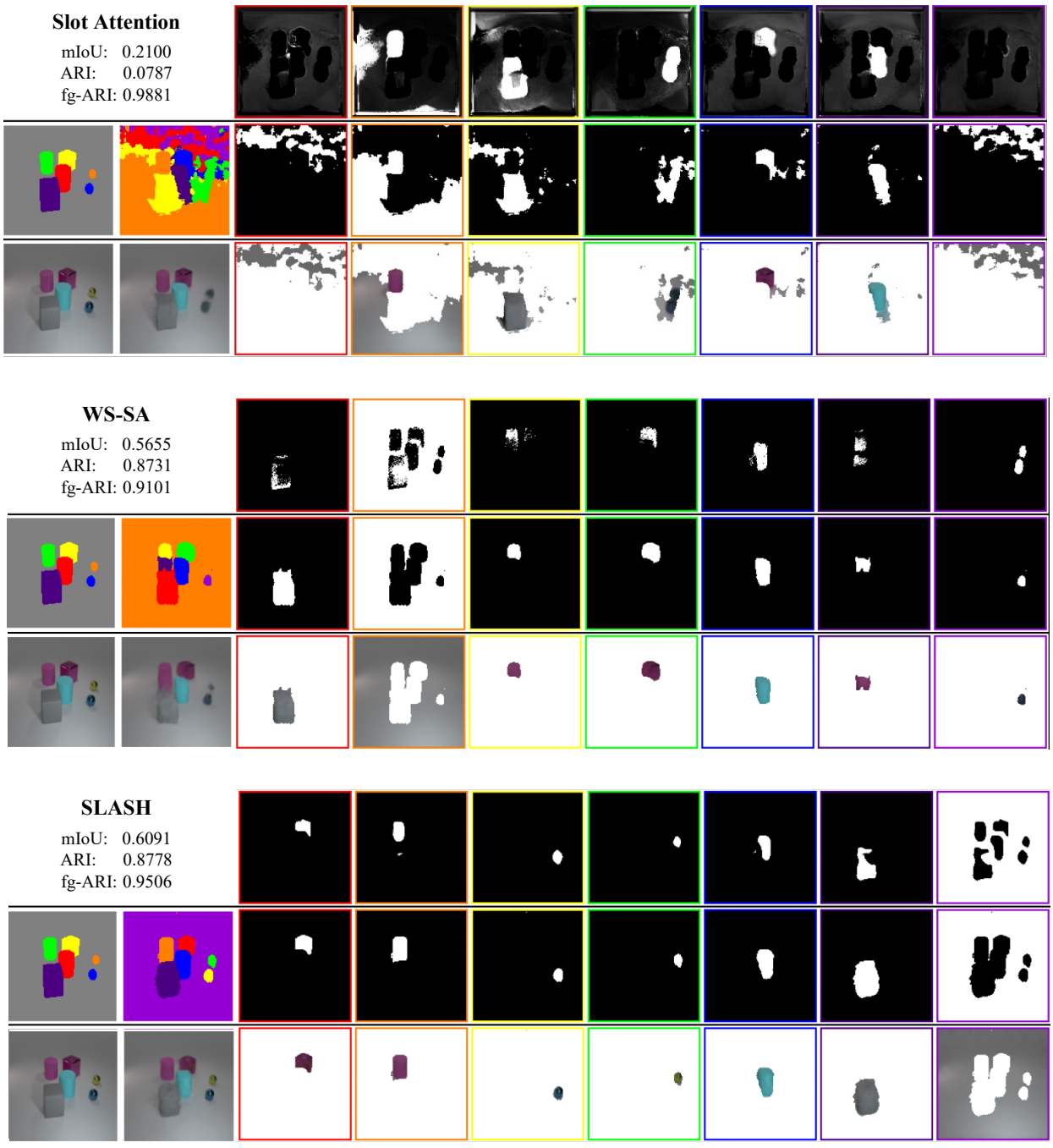

Figure 6. Results on CLEVR6 by the second-worst models from SA, WS-SA and SLASH. You can find the figure details in Sec. 4.2.

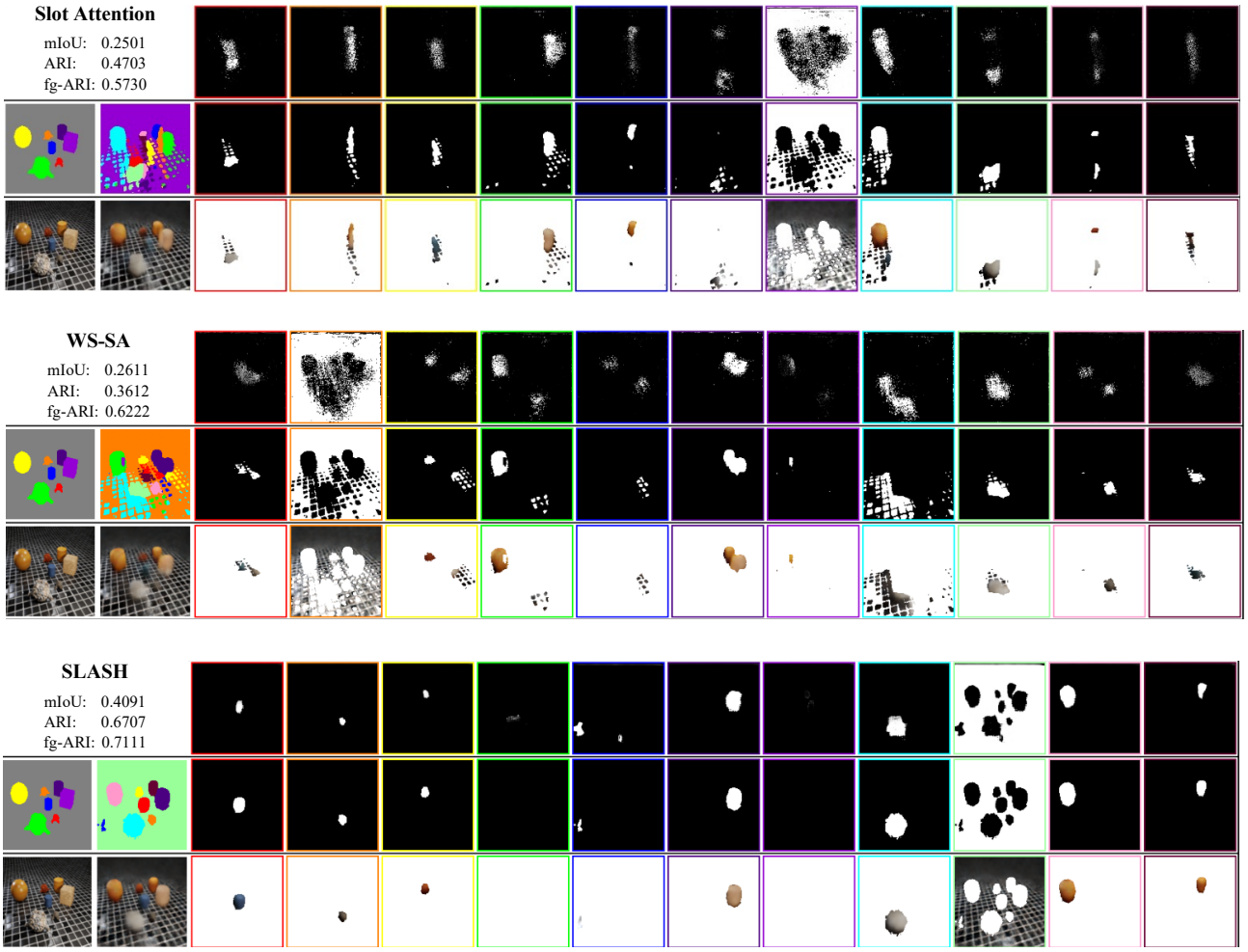

Figure 7. Results on CLEVRTEX by the second-best models from SA, WS-SA and SLASH. You can find the figure details in Sec. 4.2.

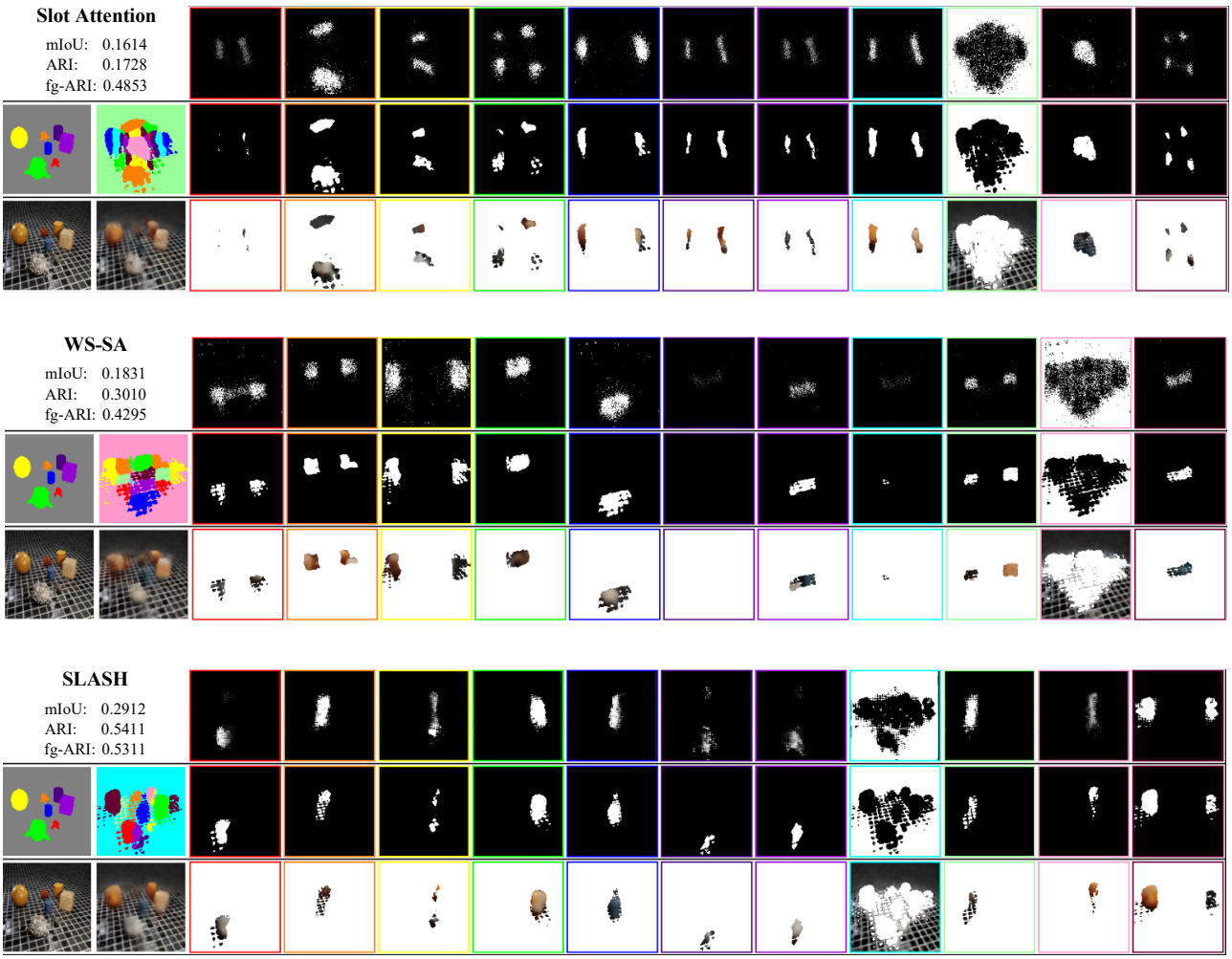

Figure 8. Results on CLEVRTEX by the second-worst models from SA, WS-SA and SLASH. You can find the figure details in Sec. 4.2.

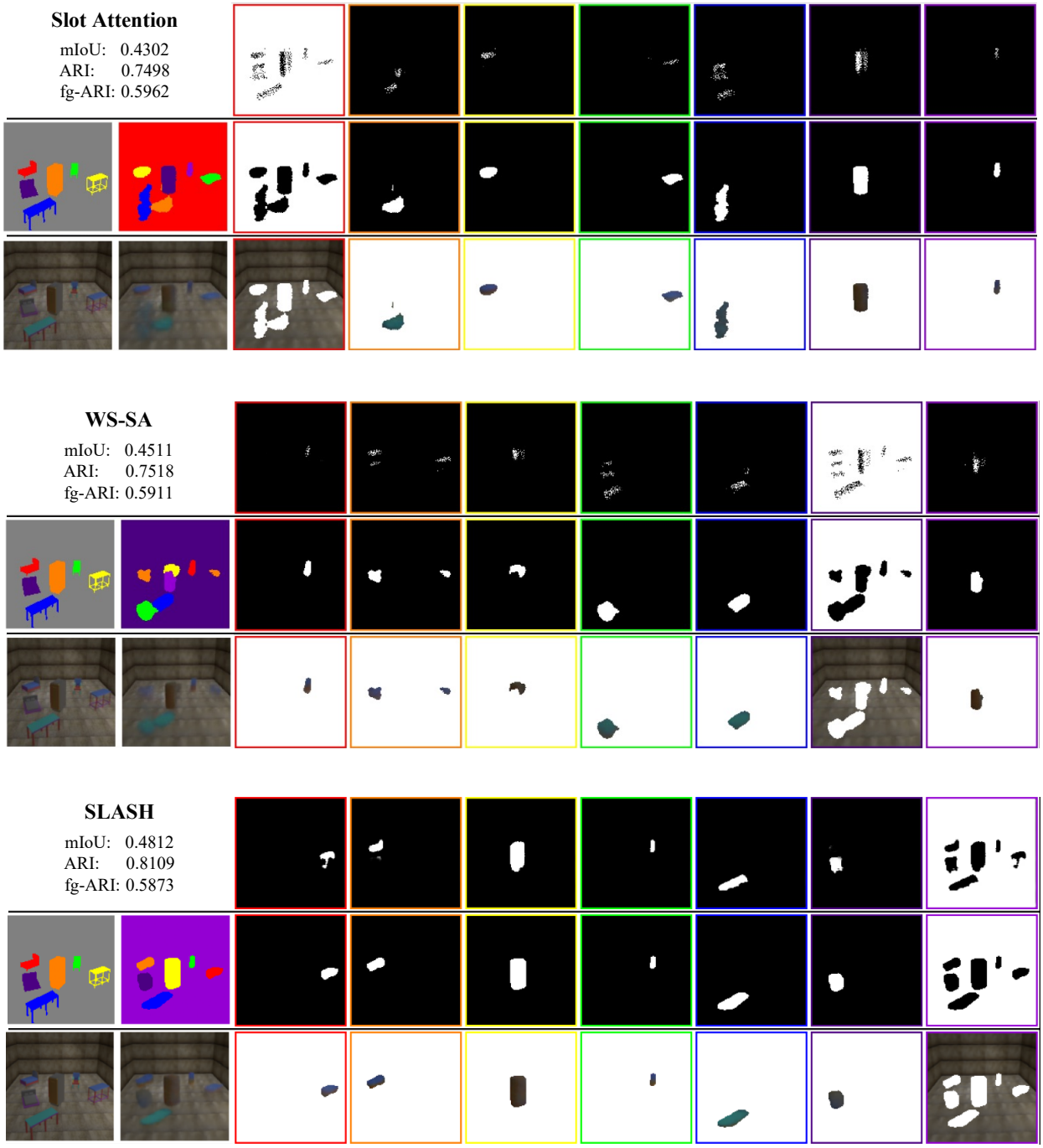

Figure 9. Results on PTR by the second-best models from SA, WS-SA and SLASH. You can find the figure details in Sec. 4.2.

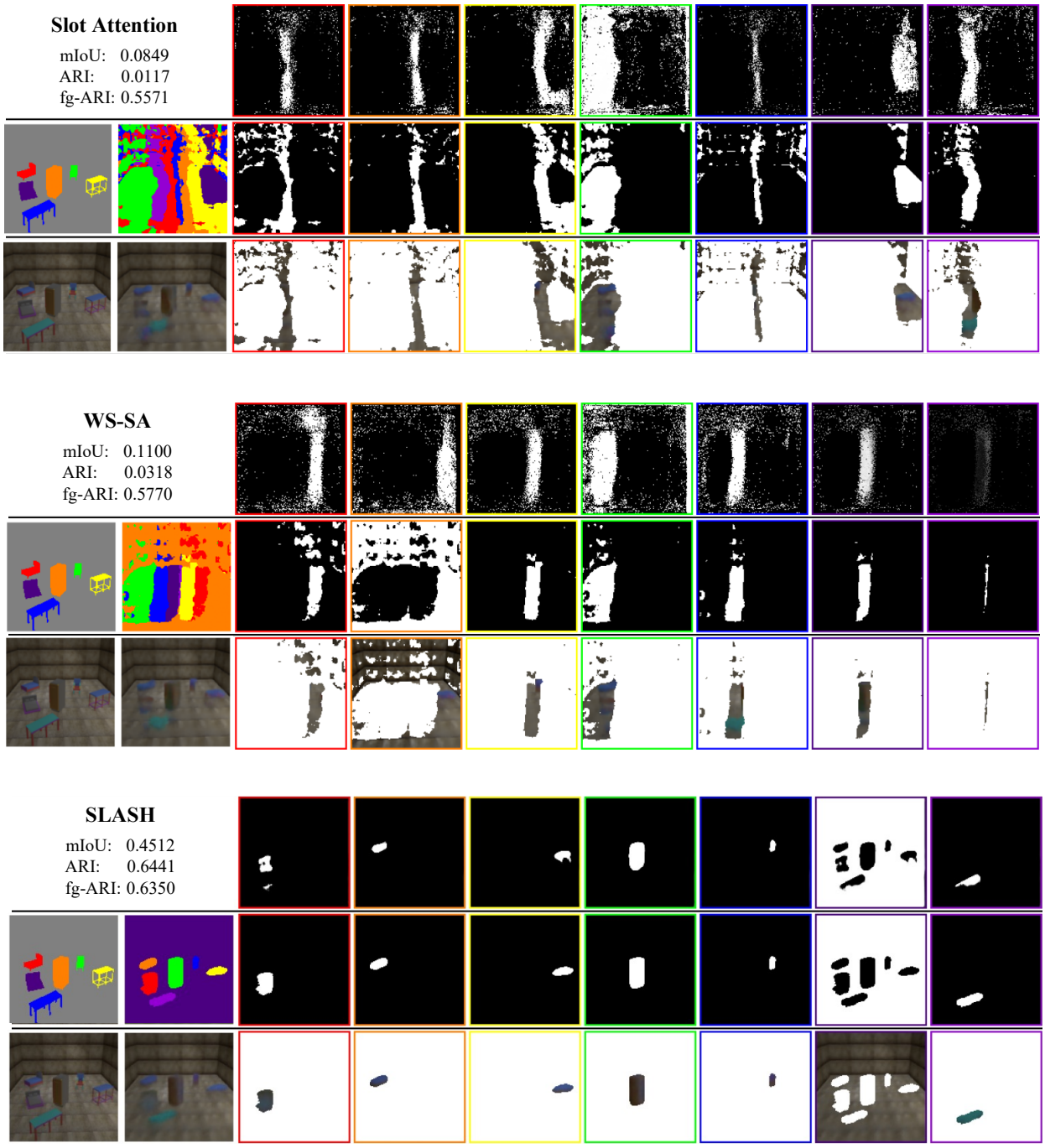

Figure 10. Results on PTR by the second-worst models from SA, WS-SA and SLASH. You can find the figure details in Sec. 4.2.

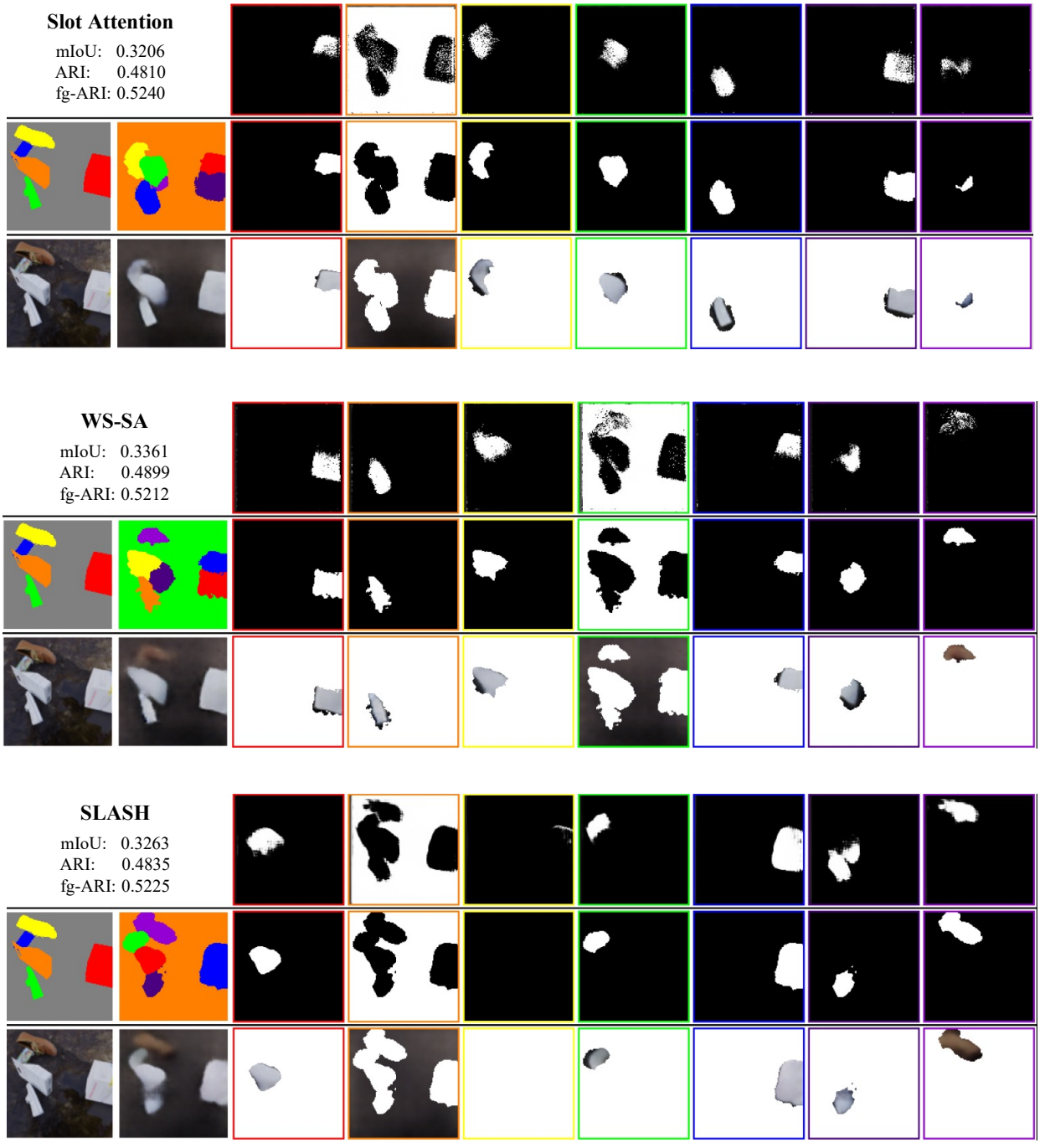

Figure 11. Results on MOVi by the second-best models from SA, WS-SA and SLASH. You can find the figure details in Sec. 4.2.

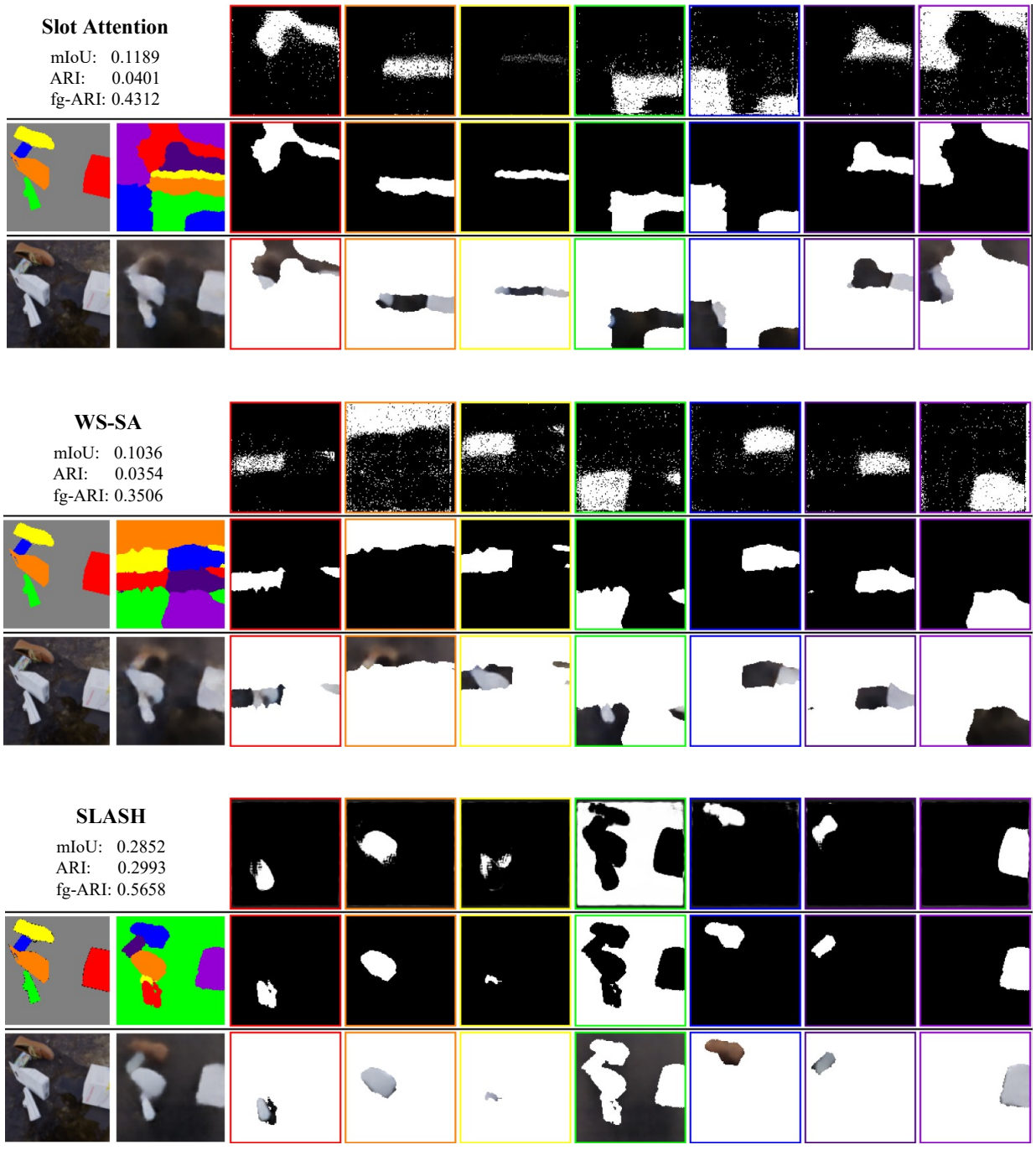

Figure 12. Results on MOVi by the second-worst models from SA, WS-SA and SLASH. You can find the figure details in Sec. 4.2.

References

[1] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 1

[2] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019. 1

[3] Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Genesis-v2: Inferring unordered object representations without iterative refinement. In *NeurIPS*, 2021. 1

[4] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. In *CVPR*, 2022. 3

[5] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *ICML*. PMLR, 2019. 1

[6] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, (50), 2020. 1

[7] Yining Hong, Li Yi, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. In *NeurIPS*, 2021. 3

[8] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 3

[9] Rishabh Kabra, Chris Burgess, Loic Matthey, Raphael Lopez Kaufman, Klaus Greff, Malcolm Reynolds, and Alexander Lerchner. Multi-object datasets. https://github.com/deepmind/multi-object-datasets/, 2019. 3

[10] Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. In *NeurIPS*, 2021. 1, 3

[11] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020. 1, 2, 3

[12] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*. PMLR, 2020. 1

[13] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022. 1

[14] Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. In *ICLR*, 2021. 1

[15] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *arXiv preprint arXiv:2205.14065*, 2022. 1

[16] A_P Varga and Roger K Moore. Hidden markov model decomposition of speech and noise. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1990. 1

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1