
Seg-VAR: Image Segmentation with Visual Autoregressive Modeling

Anonymous Author(s)

Affiliation

Address

email

1 This supplementary material provides more details about the proposed Seg-VAR. The first part
2 includes discussions about the design of Seg-VAR and its comparison with previous methods,
3 followed by the implementation details. Then, we provide extra ablation experiments. What’s more,
4 we include the visualizations of our model. The content is organized as follows:

- 5 • Discussions of Seg-VAR’s differences with previous methods.
- 6 • The implementation details of Seg-VAR.
- 7 • More ablation study experiment of Seg-VAR.
- 8 • Visualizations of Seg-VAR.

9 1 Discussion

10 Seg-VAR is a novel framework that rethinks segmentation as a conditional autoregressive mask
11 generation problem. Compared to previous works (discriminative modeling and diffusion-based seg-
12 mentation), our Seg-VAR has several advantages: **1) Paradigm Shift in Task Definition.** Traditional
13 segmentation approaches (e.g., Mask2Former, Mask R-CNN) treat segmentation as a discriminative
14 pixel-classification task. In contrast, Seg-VAR fundamentally redefines segmentation as a conditional
15 autoregressive generative process through latent space modeling. This paradigm shift enables three
16 critical advantages: *Hierarchical Spatial Reasoning*: Our sequential mask generation mimics human
17 visual perception by progressively refining object boundaries and instance relationships. *Unified*
18 *Representation*: The seglat encoding unifies semantic/instance/panoptic segmentation through a
19 shared latent language, eliminating task-specific architectural modifications. *Error Propagation*
20 *Resilience*: Unlike discriminative models, where local misclassifications corrupt global outputs, our
21 autoregressive mechanism allows self-correcting predictions through sequential dependency modeling.
22 **2) Latent Space Innovation.** The proposed spatially-aware seglat encoding addresses the limitation
23 of generative segmentation: *Instance Ambiguity Resolution*: Conventional latent representations (e.g.,
24 VQ-VAE) fail to distinguish overlapping instances. Our location-sensitive color mapping injects
25 spatial coordinates into tokenization, achieving higher instance separation accuracy in COCO val2017
26 compared to vanilla VQ encoding. **3) Unified Generation & Perception System.** The autoregressive
27 formulation of Seg-VAR not only advances segmentation performance but also lays the groundwork
28 for unifying visual perception and generation—a long-standing challenge in computer vision. This
29 direction suggests a future where segmentation and generation are not isolated tasks but different
30 operational modes of a single autoregressive engine.

31 2 Implementation Details

32 **Panoptic and instance segmentation.** We operate all experiments with 8 V100 GPUs. We use
33 Detectron2 [11] and follow the updated Mask R-CNN [7] baseline settings for the COCO dataset.
34 More specifically, we use AdamW [9] optimizer and the step learning rate schedule. We use an initial
35 learning rate of 0.0001 and a weight decay of 0.05 for all backbones. A learning rate multiplier of 0.1

is applied to the backbone and we decay the learning rate at 0.9 and 0.95 fractions of the total number of training steps by a factor of 10. Training iterations are also reported in all experimental figures. For data augmentation, we use the large-scale jittering (LSJ) augmentation [6, 5] with a random scale sampled from the range 0.1 to 2.0 followed by a fixed size crop to 1024×1024 . We use the standard Mask R-CNN inference setting where we resize an image with shorter side to 800 and longer side up-to 1333. We also report FLOPs and fps. FLOPs are averaged over 100 validation images (COCO images have varying sizes). Frames-per-second (fps) is measured on a V100 GPU with a batch size of 1 by taking the average runtime on the entire validation set including post-processing time.

Semantic segmentation. We follow the same settings as [2] to train our models, except: 1) a learning rate multiplier of 0.1 is applied to *both* CNN and Transformer backbones instead of only applying it to CNN backbones in [3], 2) both ResNet and Swin backbones use an initial learning rate of 0.0001 and a weight decay of 0.05, instead of using different learning rates in [3].

VAR modeling. We follow VAR [10] and ControlVAR [8]. During training, we leverage the pre-trained VAR tokenizer to tokenize seglat and control. The training details follow the strategy in ControlVAR. For each depth, we train the model for 30 epochs with an Adam optimizer. We follow the same learning rate and weight decay as VAR. To apply the classifier-free guidance, we replace class and control type conditions with empty tokens with 0.1 probability. For inference, we utilize top-k top-p sampling with $k=900$ and $p=0.96$ for encoding and decoding the seglat.

3 Additional Ablation Experiment

Additional video segmentation benchmarks. As shown in Table. 1, we conduct additional experiments on image semantic, instance, and panoptic segmentation. As illustrated in the table, our Seg-VAR has surpassed GSS across all metrics and all segmentation tasks by a large margin. Even though GSS is adapted with positional encoding to accommodate to various segmentation purposes, its performance is still lower than our novel unified generative design.

Method	ADE20K [12]		CityScapes [4]		
	AP	mIoU	AP	mIoU	PQ
GSS [†] [1]	36.3	48.5	43.1	80.1	59.3
Seg-VAR	43.2	54.9	49.5	85.8	66.8

Table 1: **Results on Image Segmentation Benchmarks.** Our Seg-VAR demonstrates better performance than GSS across different benchmarks and tasks.

Image-control Generation. We conduct an experiment on the image FID comparison of different models in Table. 2. While our model exhibits a marginal performance decrease compared to VAR (likely attributable to the added complexity of integrating control mechanisms), it consistently outperforms ControlVAR. Notably, the performance gap diminishes with increasing model scale, suggesting that effectively modeling both image content and control signals demands greater network capacity than image-only modeling.

Depth	16	20	24	30
VAR [10]	3.60	2.95	2.33	1.97
ControlVAR [8]	4.25	3.25	2.69	1.98
Seg-VAR	3.8	3.05	2.56	1.97

Table 2: **Image FID Comparison.** Our Seg-VAR demonstrates better performance than ControlVAR.

Architecture Designs. As shown in Table. 3, we experimented different designs of encoding and decoding seglats. As illustrated in the table, ‘UU’ settings only outperforms GSS by a smaller margin, while the final model greatly improves the performance (6.9 AP on ADE20K, for example). This indicates the significance of jointly training seglat encoder and decoder.

Method	ADE20K		CityScapes	
	AP	mIoU	AP	mIoU
GSS [†] [1]	36.3	48.5	43.1	80.1
Seg-VAR-UU	38.6	50.4	44.8	81.3
Seg-VAR-TU	41.3	53.0	47.8	84.0
Seg-VAR	43.2	54.9	49.5	85.8

Table 3: **Different encoder designs.** ‘UU’ indicates using untrained VAR as seglat encoder and decoder, while ‘Tu’ indicates we finetune the encoder. The performance indicates the sigficance of jointly training with seglat encoder and decoder.

4 Visualization

As shown in Figure. 1, we provide visualizations of Seg-VAR’s outputs for the video temporal grounding, video QA, and video reasoning segmentation. Compared to the SFT-based method, our Seg-VAR demonstrates better results in multiple video perception and understanding tasks, indicating that by applying spatiotemporal aggregated reinforcement, our method is capable of greatly improving video perception capacity from different perspectives.



Figure 1: **Image Instance Segmentation Examples.** The examples show that Seg-VAR can successfully discriminate multiple instances in the crowded scene. While other methods fail to segment and classify these instances properly. GSS* indicated that we add positional encoding to the method.

References

- [1] Jiaqi Chen, Jiachen Lu, Xiatian Zhu, and Li Zhang. Generative semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7111–7120, 2023. 2, 3
- [2] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv:2112.10764*, 2021. 2
- [3] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 2
- [4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2
- [5] Xianzhi Du, Barret Zoph, Wei-Chih Hung, and Tsung-Yi Lin. Simple training strategies and model scaling for object detection. *arXiv preprint arXiv:2107.00057*, 2021. 2
- [6] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 2
- [7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1
- [8] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024. 2
- [9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1
- [10] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 2
- [11] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 1
- [12] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 2