Diffusion Action Segmentation: Supplementary

Daochang Liu[1], Qiyue Li[2,4], Anh-Dung Dinh[1], Tingting Jiang[2], Mubarak Shah[3], Chang Xu[1]
[1]School of Computer Science, Faculty of Engineering, The University of Sydney
[2]NERCVT, NKLMIP, School of Computer Science, [4]School of Mathematical Sciences, Peking University
[3]Center for Research in Computer Vision, University of Central Florida
{daochang.liu, c.xu}@sydney.edu.au shah@crcv.ucf.edu

This supplementary material includes more implementation details, experimental comparison, qualitative results, and some other early attempts we consider interesting.

1. More Implementation Details

For the Gaussian smoothing when obtaining the soft ground truth of action boundaries $\bar{B} = \lambda(B)$, the standard deviation of the Gaussian kernel is set as 1, 20, 3 for GTEA, 50Salads, and Breakfast respectively. This is consistent with the different video lengths in different datasets. We set these Gaussian kernels to make $\bar{B}$ have similar bell curve shapes across the three datasets. For the decoder, the step embedding is of 512 dimensions. When using the re-implemented ASFormer [10] decoder as our decoder g_ψ, the concatenation of the conditioning features $E \odot M$ (or E at inference) and the noisy sequence Y_s (or $\hat{Y}_s$ at inference) is used as queries and keys in the cross-attention, while the noisy sequence Y_s (or $\hat{Y}_s$ at inference) is taken as values. The step embedding is added to the values. When using the single-stage model in MS-TCN [4] as our decoder g_ψ, the concatenation of the conditioning features $E \odot M$ (or E at inference) and the noisy sequence Y_s (or $\hat{Y}_s$ at inference) is used as the input. The step embedding is added to the input. Our method does not use positional encoding since it was found harmful in the original ASFormer paper [10]. Our model can be trained on a single NVIDIA RTX 2080 GPU.

2. Other Early Attempts

In this section, we would like to share with the readers several preliminary attempts made at the early stage of this research, which are immature, not benchmarked, but might be inspirational.

Different Forms of Condition Masking. Human actions are predictable to some degree if we observe what has happened in the past. Therefore, we tried to mask the conditioning features after a random time location to enhance the future predictive learning of the model. Similarly, we also tried an inverted way by masking past features. Another form we attempted was a fully random mask that blocks random short clips in the video. These forms mentioned above were not evidently helpful in our preliminary experiments. But it is promising to explore more potential forms in the future given the flexibility of our condition masking strategy.

Combining Masking Schemes at Inference. Our method uses no masking for the conditioning features at the inference time. We also tried to infer differently. As discussed in the main paper, our explicit prior modeling can be interpreted from the perspective in the classifier-free guidance of the diffusion model [5]. We can regard the model with no masking ($M^{\mathtt{N}}$) as a fully conditional generation and the model with all masking ($M^{\mathtt{P}}$) as an unconditional generation. The classifier-free guidance combines a conditional diffusion model and an unconditional diffusion model by a weighted aggregation of their outputs at each update step to improve the generation quality. Therefore, we tried to aggregate the outputs using $M^{\mathtt{N}}$ and $M^{\mathtt{P}}$ at each inference step. In our early experiments, it was noticed that the model could achieve a better edit score if we put higher weights on the outputs using $M^{\mathtt{P}}$, and a better accuracy if we put higher on the outputs using $M^{\mathtt{N}}$, but not both at the same time. We suspected this is because of the interruptive predictions at boundaries when using $M^{\mathtt{N}}$. Then we tried to apply an adaptive boundary-aware approach for the aggregation weights that puts smaller weights at boundaries for the outputs using $M^{\mathtt{N}}$. However, we found it non-trivial to reliably detect the boundaries at the inference.

Further explorations beyond these early attempts are possible based on our extendable framework.

3. Comparison with Methods on arXiv

Table 1 provides a comparison between our method and several recent methods on arXiv. This comparison does not change the conclusion in the main paper.

Method	GTEA F1@{10, 25, 50}	 Edit	 Acc	 Avg	50Salads F1@{10, 25, 50}	 Edit	 Acc	 Avg	Breakfast F1@{10, 25, 50}	 Edit	 Acc	 Avg
[7]C2F-TCN, *arXiv'21*	90.3 / 88.8 / 77.7	86.4	80.8	84.8	84.3 / 81.8 / 72.6	76.4	84.9	80.0	72.2 / 68.7 / 57.6	69.6	76.0	68.8
[8]CETNet, *arXiv'22*	91.8 / 91.2 / 81.3	87.9	80.3	86.5	87.6 / 86.5 / 80.1	81.7	86.9	84.6	79.3 / 74.3 / 61.9	77.8	74.9	73.6
[3]TUT, *arXiv'22*	89.0 / 86.4 / 73.3	84.1	76.1	81.8	89.3 / 88.3 / 81.7	84.0	87.2	86.1	76.2 / 71.9 / 60.0	73.7	76.0	71.6
[6]Liu *et al.*, *arXiv'23*	91.4 / 90.2 / 82.1	86.6	80.3	86.1	87.9 / 86.6 / 80.5	82.7	86.6	84.9	77.5 / 72.3 / 59.5	76.7	73.7	71.9
[11]S2G, *arXiv'22*	95.7 / 94.2 / 91.3	92.0	89.8	92.6	91.5 / 90.2 / 87.3	89.1	88.6	89.3	- / - / -	-	-	-
***DiffAct**, Ours*	**92.5 / 91.5 / 84.7**	**89.6**	**82.2**	**88.1**	**90.1 / 89.2 / 83.7**	**85.0**	**88.9**	**87.4**	**80.3 / 75.9 / 64.6**	**78.4**	**76.4**	**75.1**

Table 1. Comparison with recent methods on arXiv. The method in gray is not suitable for direct comparison due to the extra usage of multi-modal features [11]. We list it here for readers' reference. This comparison does not change the conclusion in the main paper.

	F1@{10, 25, 50}	Edit	Acc	Avg
Macro Mean	90.3 / 89.4 / 83.9	85.0	88.8	87.5
Macro Std.	0.08 / 0.09 / 0.14	0.16	0.03	0.08
Micro Mean	90.3 / 89.4 / 83.9	85.0	88.7	87.4
Micro Std.	0.86 / 1.03 / 1.25	1.17	0.36	0.93

Table 2. Inference stability on 50Salads

	F1@{10, 25, 50}	Edit	Acc	Avg
Mean	90.4 / 89.4 / 83.7	84.9	88.5	87.4
Std.	0.11 / 0.12 / 0.18	0.18	0.16	0.15

Table 3. Training stability on 50Salads

	F1@{10, 25, 50}	Edit	Acc	Avg
Length Top 50%	88.9 / 88.2 / 82.1	83.4	88.9	86.3
Length Bottom 50%	91.8 / 90.8 / 85.9	87.2	89.0	89.0
#Actions Top 50%	88.3 / 87.0 / 81.6	81.2	88.6	85.3
#Actions Bottom 50%	91.7 / 91.0 / 85.1	87.5	88.7	88.8

Table 4. Effects of video length and action number on 50Salads

	F1@{10, 25, 50}	Edit	Acc	Avg
Baseline adapted from [1, 9]	63.2 / 60.3 / 51.7	52.6	81.0	61.8
***DiffAct**, Ours*	90.1 / 89.2 / 83.7	85.0	88.9	87.4

Table 5. Diffusion image segmentation baseline on 50Salads

4. Stability

Inference Stability. We fix the seed at inference for all experiments in the paper to remove inference randomness. Here we further provide both macro and micro inference stability on 50Salads. The *macro* setting follows the evaluation convention, which repeats the experiment as a whole with ten different seeds at inference. The *micro* setting repeats the inference for each video with ten different seeds and averages the deviations over videos. The macro result should be used when comparing to the state-of-the-art. Our method is highly stable at inference as in Table 2.

Training Stability. We re-run the main experiment on 50Salads ten times with different training seeds and the same inference seed for a training stability check. The mean values and the standard deviations are reported in Table 3, from which we can see the results are stable with narrow deviations.

5. Effects of Video Length and Action Number

We report results on 50Salads in Table 4 by dividing test videos into top and bottom halves to investigate the impact of the video length and action number. Our method performs well regardless of these factors.

6. Discussion

Diffusion models have been employed for image segmentation [2, 1, 9]. Our diffusion model for video understanding differs from diffusion-based image segmentation, by customizing the diffusion pipeline and introducing unique prior modeling for action analysis. Diffusion image segmentation, *e.g.*, SegDiff [1] and MedSegDiff [9] was built on U-Net with an objective of noise ϵ prediction measured by vanilla L2 loss. In contrast, we adapt ASFormer and suggest x_0 prediction as a more appropriate objective for our task, and cross-entropy loss, smoothness loss, and boundary loss are investigated together for a comprehensive objective. As in Table 5, a naive application of diffusion image segmentation by simply changing the data modality to video results in a much lower performance.

7. More Qualitative Results

This section presents more qualitative results from Fig. 4 to Fig. 21. The predictions and the ground truth sequences are visualized for randomly selected videos from the GTEA, 50Salads, and Breakfast datasets. Different datasets use different sets of color codes in the plots. In general, our model can achieve accurate and temporally coherent results and excellent overall performance.

8. More Results using M^{P} at Inference

To explore situations of unconditional generation, we provide more results using M^{P} at inference in Fig. 1, Fig. 2, and Fig. 3 for the three datasets respectively. Different datasets use different sets of color codes in the plots. These results show that our model is able to generate ***broadly plausible*** action sequences even when all the conditions are masked. It is interesting that the generated action sequences

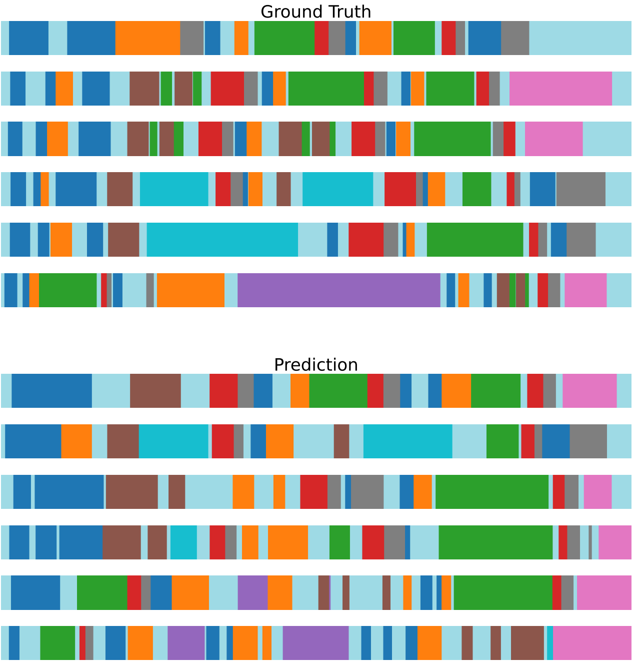

Figure 1. **Top**: Ground truth action sequences from GTEA. **Bottom**: Results using M^{P} at inference for GTEA. Our model is able to capture the distribution of actions and generate sequences roughly similar to real sequences when all conditions are masked.

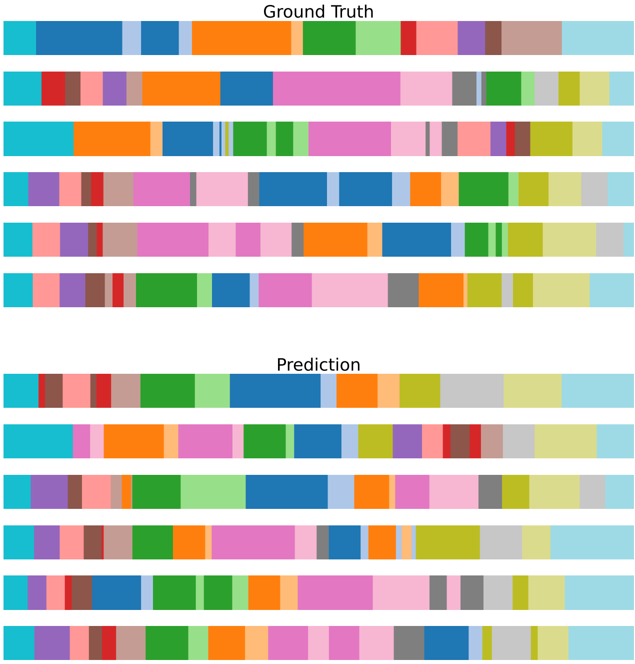

Figure 2. **Top**: Ground truth action sequences from 50Salads. **Bottom**: Results using M^{P} at inference for 50Salads. Our model is able to capture the distribution of actions and generate sequences roughly similar to real sequences when all conditions are masked.

exhibit the characteristics of each dataset. This validates our model's ability in capturing the prior distributions of action sequences via generative learning.

References

[1] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. SegDiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 2

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *ICLR*, 2021. 2

[3] Dazhao Du, Bing Su, Yu Li, Zhongang Qi, Lingyu Si, and Ying Shan. Do we really need temporal convolutions in action segmentation? *arXiv preprint arXiv:2205.13425*, 2022. 2

[4] Yazan Abu Farha and Jurgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, 2019. 1

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[6] Zhichao Liu, Leshan Wang, Desen Zhou, Jian Wang, Songyang Zhang, Yang Bai, Errui Ding, and Rui Fan. Temporal segment transformer for action segmentation. *arXiv preprint arXiv:2302.13074*, 2023. 2

[7] Dipika Singhania, Rahul Rahaman, and Angela Yao. Coarse to fine multi-resolution temporal convolutional network. *arXiv preprint arXiv:2105.10859*, 2021. 2

[8] Jiahui Wang, Zhenyou Wang, Shanna Zhuang, and Hui Wang. Cross-enhancement transformer for action segmentation. *arXiv preprint arXiv:2205.09445*, 2022. 2

[9] Junde Wu, Huihui Fang, Yu Zhang, Yehui Yang, and Yanwu Xu. MedSegDiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint arXiv:2211.00611*, 2022. 2

[10] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. ASFormer: Transformer for action segmentation. In *BMVC*, 2021. 1

[11] Junbin Zhang, Pei-Hsuan Tsai, and Meng-Hsun Tsai. Semantic2graph: Graph-based multi-modal feature for action segmentation in videos. *arXiv preprint arXiv:2209.05653*, 2022. 2

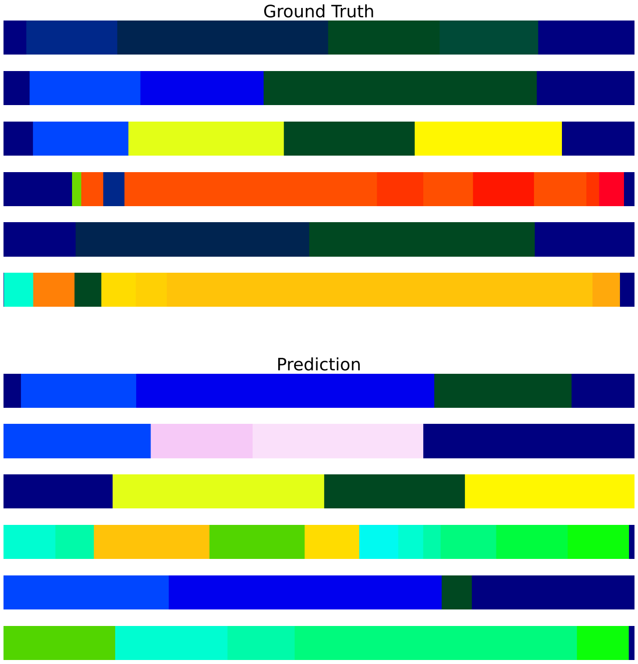

Figure 3. **Top**: Ground truth action sequences from Breakfast. **Bottom**: Results using M^{P} at inference for Breakfast. Our model is able to capture the distribution of actions and generate sequences roughly similar to real sequences when all conditions are masked. Note that Breakfast tends to have very distinct sets of action classes across videos.

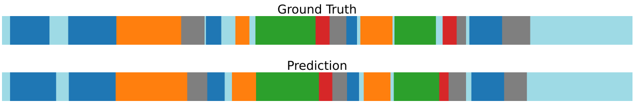

Figure 4. Video ‘S1_Cheese_C1’ from GTEA.

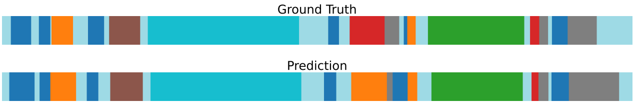

Figure 5. Video ‘S1_Peanut_C1’ from GTEA.

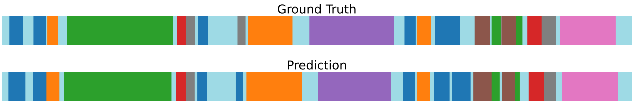

Figure 6. Video ‘S2_Tea_C1’ from GTEA.

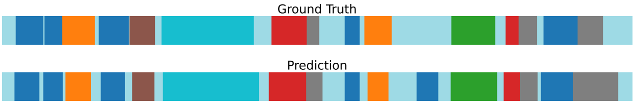

Figure 7. Video ‘S3_Peanut_C1’ from GTEA.

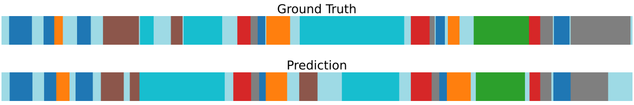

Figure 8. Video ‘S4_Pealate_C1’ from GTEA.

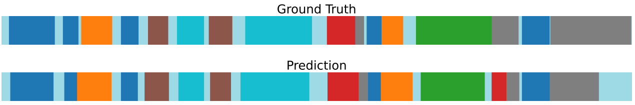

Figure 9. Video ‘S4_Peanut_C1’ from GTEA.

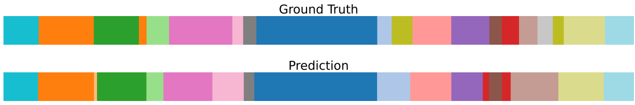

Figure 10. Video ‘rgb-09-1’ from 50Salads.

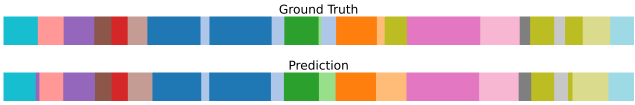

Figure 11. Video ‘rgb-05-2’ from 50Salads.

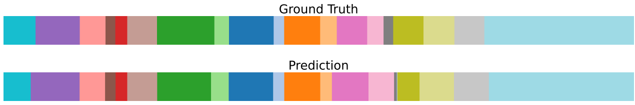

Figure 12. Video ‘rgb-10-2’ from 50Salads.

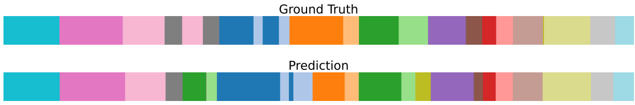

Figure 13. Video ‘rgb-15-1’ from 50Salads.

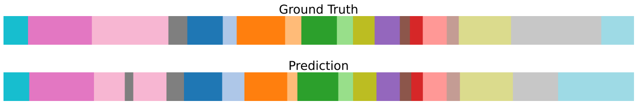

Figure 14. Video ‘rgb-17-2’ from 50Salads.

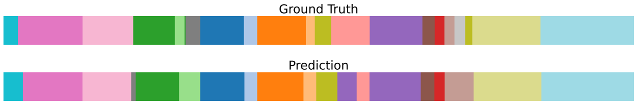

Figure 15. Video ‘rgb-24-1’ from 50Salads.

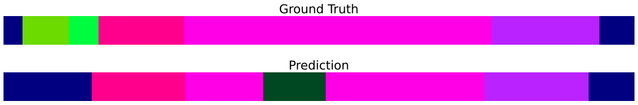

Figure 16. Video ‘P21_webcam02_P21_sandwich’ from Breakfast.

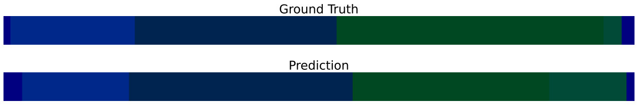

Figure 17. Video ‘P05_cam01_P05_cereals’ from Breakfast.

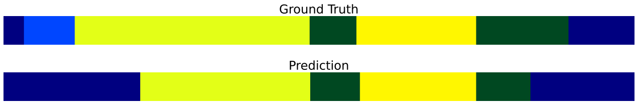

Figure 18. Video ‘P05_stereo01_P05_milk’ from Breakfast.

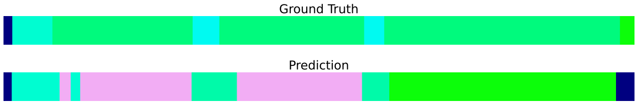

Figure 19. Video ‘P09_cam01_P09_friedegg’ from Breakfast.

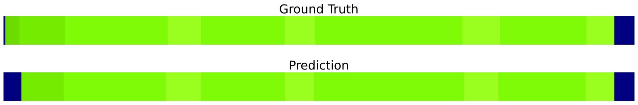

Figure 20. Video ‘P16_stereo01_P16_juice’ from Breakfast.

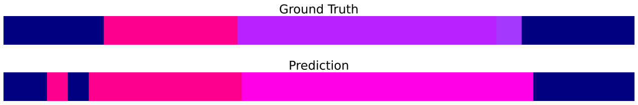

Figure 21. Video ‘P17_cam01_P17_sandwich’ from Breakfast.