MANTA: Diffusion Mamba for Efficient and Effective Stochastic Long-Term Dense Action Anticipation

Supplementary Material

Here, we present additional dataset and implementation details, as well as additional quantitative and qualitative results for our proposed MANTA model. More precisely, we discuss the implementation of our model in Sec. 1 and provide additional details about the utilized datasets in Sec. 2. Next, in Sec. 3 we present additional ablation studies for MANTA. Finally, in Sec. 4 we present additional qualitative comparisons of MANTA to previous work.

1. Implementation details

We implemented our model using Pytorch. As per Tab. 3, we use a total of $B = 15$ MANTA blocks for our final model. Our proposed network is trained for 90 epochs using the Adam [1] optimizer with a learning rate of 0.0005 for Breakfast and Assembly, and 0.001 for 50Salads. Following [2], we use $T = 1000$ diffusion steps for training, $D = 50$ DDIM steps for inference on Breakfast and Assembly101, and $D = 10$ inference steps for 50Salads.

2. Datasets

In Tab. 1, we show additional details for the datasets used in our work. Specifically, we provide average and maximum video durations, as well as the average and maximum number of individual segments per video. Since in our adopted anticipation protocol only up to 50% of the video frames are utilized for anticipation, we additionally provide the statistics for the corresponding intervals of the videos in brackets in blue, including only frames falling into the anticipation intervals. As one can observe, the long temporal horizon of videos used for future anticipation and the numerous action segments that need to be predicted highlight the long-term nature of the addressed task.

Dataset	Avg. Num. Seg.	Max. Num. Seg	Avg. Dur. (min)	Max. Dur. (min)
Breakfast	7 (3)	25 (15)	2.3 (1.2)	10.8 (5.4)
50Salads	20 (11)	26 (18)	6.4 (3.2)	10.1 (5.1)
Assembly101	12 (5)	73 (40)	3.5 (1.8)	25.0 (12.5)

Table 1. *(Left)* Number of segments and *(Right)* duration for the whole video and in the anticipation interval.

3. Ablation Study

3.1. Bidirectional State-Space Layer (BSSL)

In Tab. 2 and Tab. 3 of the main paper, we analyzed the *selectivity* and *bi-directionality* of the proposed BSS layer. Here, we examine how *independent forward and backward* scanning contributes to the final performance of the MANTA model. More specifically, we tested if having *independent* parameters for *forward and backward* scanning branches is the best way to structure the BSS layer. To investigate this, we evaluated the effect of weight-sharing between the two branches. As shown in Tab. 2, while Top-1 MoC accuracy is similar across networks with shared and independent weights, Mean MoC accuracy is higher in the model with branch-specific weights. We, therefore, keep the weights separate for the two BSSL branches.

MoC	Shared	β ($\alpha = 0.2$)				β ($\alpha = 0.3$)			
		0.1	*0.2*	*0.3*	*0.5*	*0.1*	*0.2*	*0.3*	*0.5*
Mean	✓	**28.5**	23.4	23.8	22.9	33.2	27.7	28.2	27.1
	✗	27.7	**25.3**	**24.6**	**23.8**	**34.2**	**30.9**	**29.1**	**27.7**
Top-1	✓	53.7	**51.0**	**48.8**	46.7	**60.7**	**55.9**	53.3	50.5
	✗	**55.5**	**51.0**	47.9	**46.9**	59.6	55.0	**53.7**	**51.9**

Table 2. Ablation of weight sharing for forward and backward branches of the BSSL on Breakfast.

3.2. MANTA Block

We experimented with varying the total number of blocks in the final model (Tab. 3). Empirically, we found that the model with $B = 15$ blocks showed the best results, with further increase or decrease in the number of blocks harming the model's performance.

MoC	Num. blocks	β ($\alpha = 0.2$)				β ($\alpha = 0.3$)			
		0.1	*0.2*	*0.3*	*0.5*	*0.1*	*0.2*	*0.3*	*0.5*
Mean	10	26.6	24.3	23.5	23.1	32.7	29.4	28.1	26.8
	15	**27.7**	**25.3**	**24.6**	**23.8**	**34.2**	**30.9**	**29.1**	**27.7**
	20	27.2	24.5	23.4	23.2	32.1	29.3	27.7	26.6
Top-1	10	54.2	49.2	46.7	46.6	57.2	52.5	52.4	49.7
	15	**55.5**	**51.0**	47.9	46.9	**59.6**	**55.0**	**53.7**	**51.9**
	20	54.7	50.5	**48.4**	**47.1**	57.8	53.4	52.3	50.5

Table 3. Ablation of the number of MANTA blocks on Breakfast.

3.3. Samples.

We analyze the effect of the total sample count on MANTA's performance in Tab. 4. The number of samples has a marginal effect on the Mean MoC accuracy whereas the Top-1 MoC increases with the number of samples. While this is expected, as Top-1 MoC only considers the sample that is closest to the ground-truth, the increase in

Top-1 MoC with a higher number of samples demonstrates the diversity of the generated predictions. Otherwise, the Top-1 MoC would saturate after a small number of samples.

MoC	Num. samples	β ($\alpha = 0.2$)				β ($\alpha = 0.3$)			
		0.1	*0.2*	*0.3*	*0.5*	*0.1*	*0.2*	*0.3*	*0.5*
Mean	5	**27.7**	**25.6**	**24.7**	**23.9**	33.9	30.7	28.9	27.5
	15	**27.7**	25.4	24.6	23.8	**34.2**	**30.9**	29.0	**27.8**
	25	25.5	25.3	24.6	23.8	**34.2**	**30.9**	**29.1**	27.7
Top-1	5	42.3	39.3	37.1	36.1	47.5	44.2	41.4	40.4
	15	51.4	47.3	44.7	44.0	55.8	52.0	50.5	48.6
	25	**55.5**	**51.0**	**47.9**	**46.9**	**59.6**	**55.0**	**53.7**	**51.9**

Table 4. Ablation of the number of samples on Breakfast.

3.4. Robustness

We report the standard deviation for the MANTA model, computed over 3 seeds on the Breakfast dataset, in Tab. 5. As expected, the std. is higher for Top-1 MoC, but the values are low, indicating the robustness of our proposed model.

MoC	β ($\alpha = 0.2$)				β ($\alpha = 0.3$)			
	0.1	*0.2*	*0.3*	*0.5*	*0.1*	*0.2*	*0.3*	*0.5*
Mean	0.07	0.07	0.06	0.05	0.04	0.18	0.09	0.11
Top-1	0.5	0.6	0.3	0.2	0.3	0.6	0.8	0.5

Table 5. Standard deviation of MANTA on Breakfast dataset computed over 3 runs with different seeds.

4. Qualitative Results

We provide qualitative comparisons of our proposed MANTA model to the previous best-performing GTDA [2] on the Breakfast dataset in Figs. 1-3, on the 50Salads dataset in Fig. 4 and on Assembly101 dataset in Fig. 5.

References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[2] Olga Zatsarynna, Emad Bahrami, Yazan Abu Farha, Gianpiero Francesca, and Juergen Gall. Gated temporal diffusion for stochastic long-term dense anticipation. *European Conference on Computer Vision (ECCV)*. 1, 2, 3, 4, 5, 6

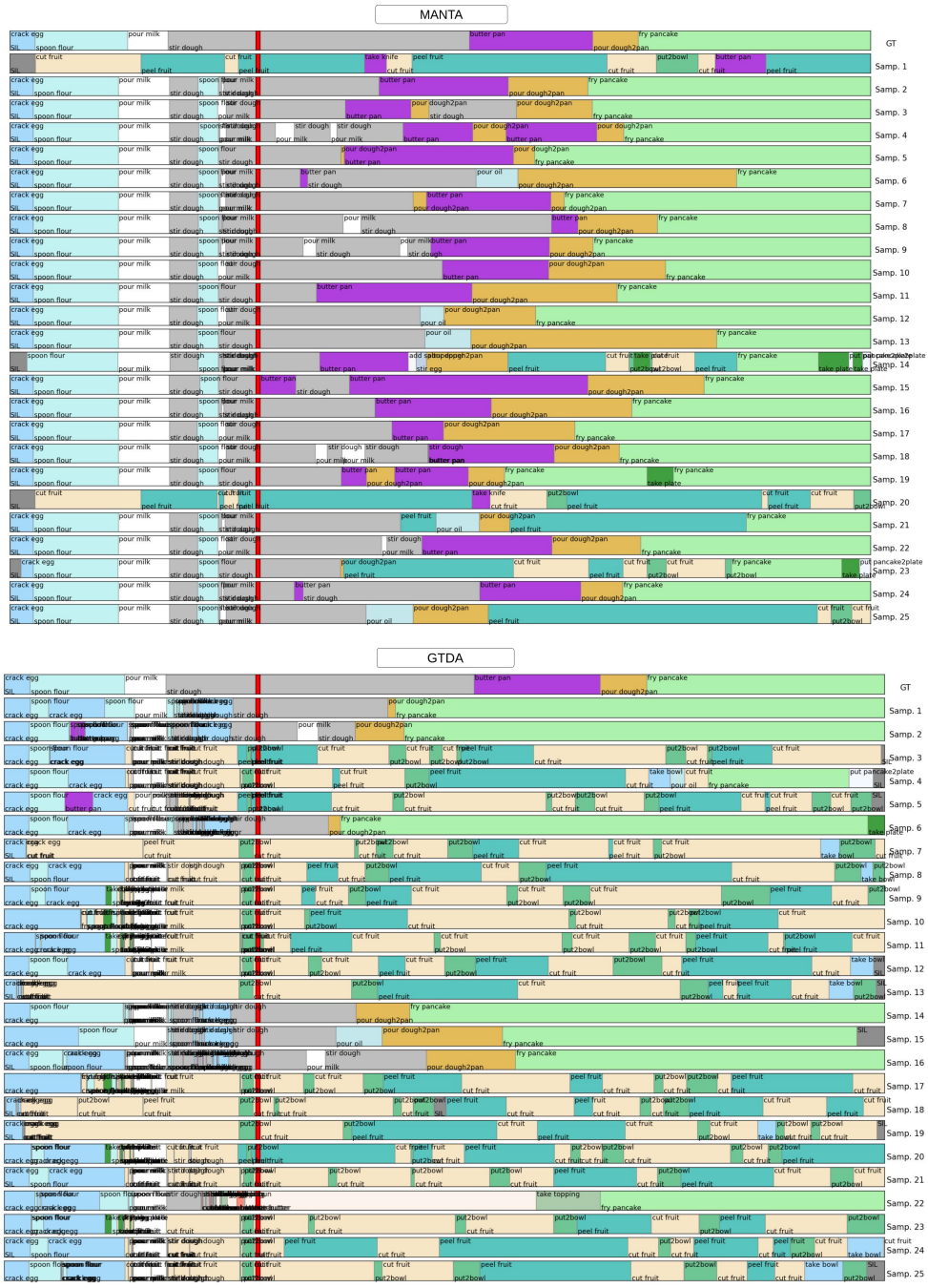

Figure 1. Qualitative comparison of MANTA *(top)* to GTDA [2] *(bottom)* on the Breakfast dataset. Best viewed zoomed in.

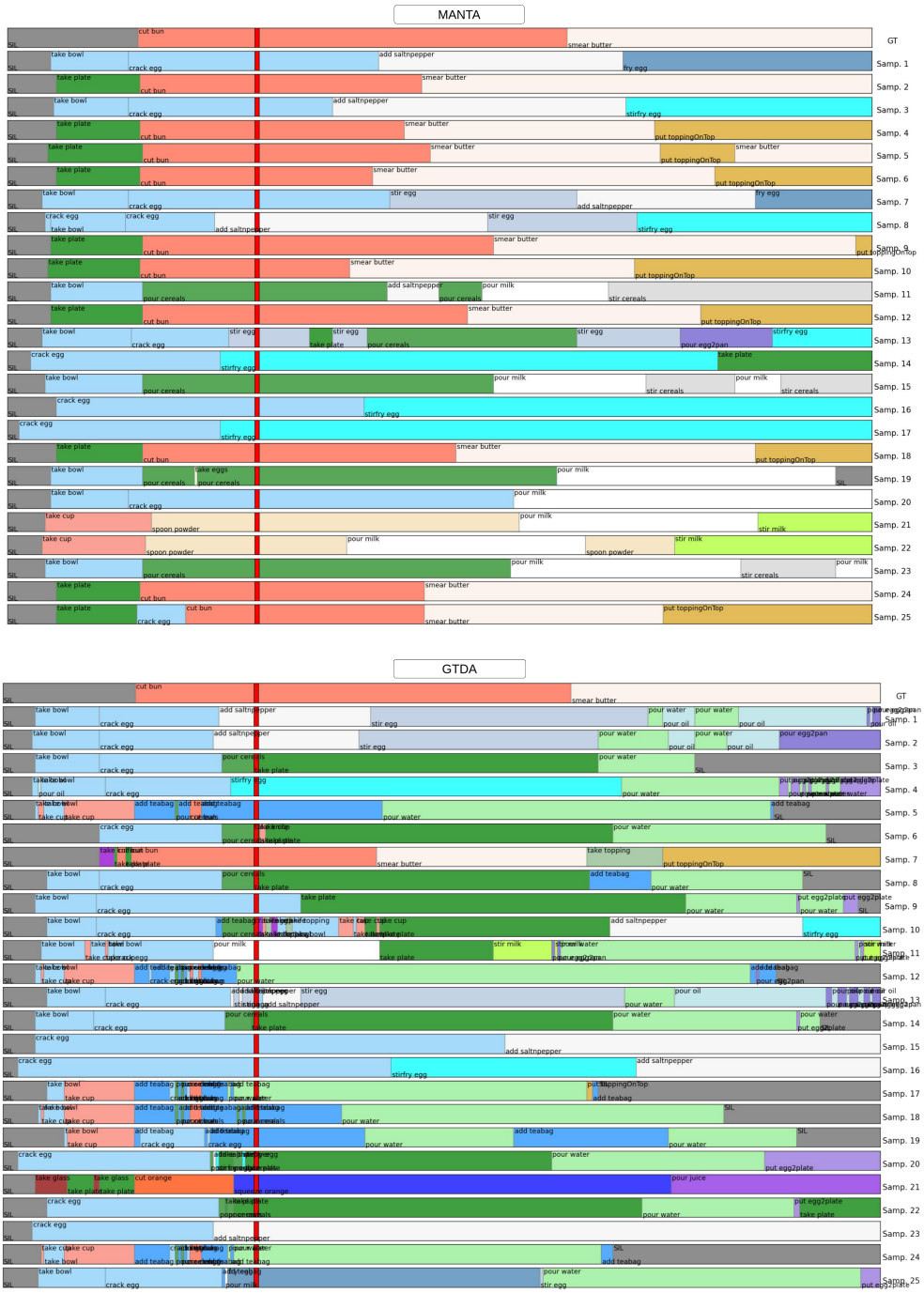

Figure 2. Qualitative comparison of MANTA *(top)* to GTDA [2] *(bottom)* on the Breakfast dataset. Best viewed zoomed in.

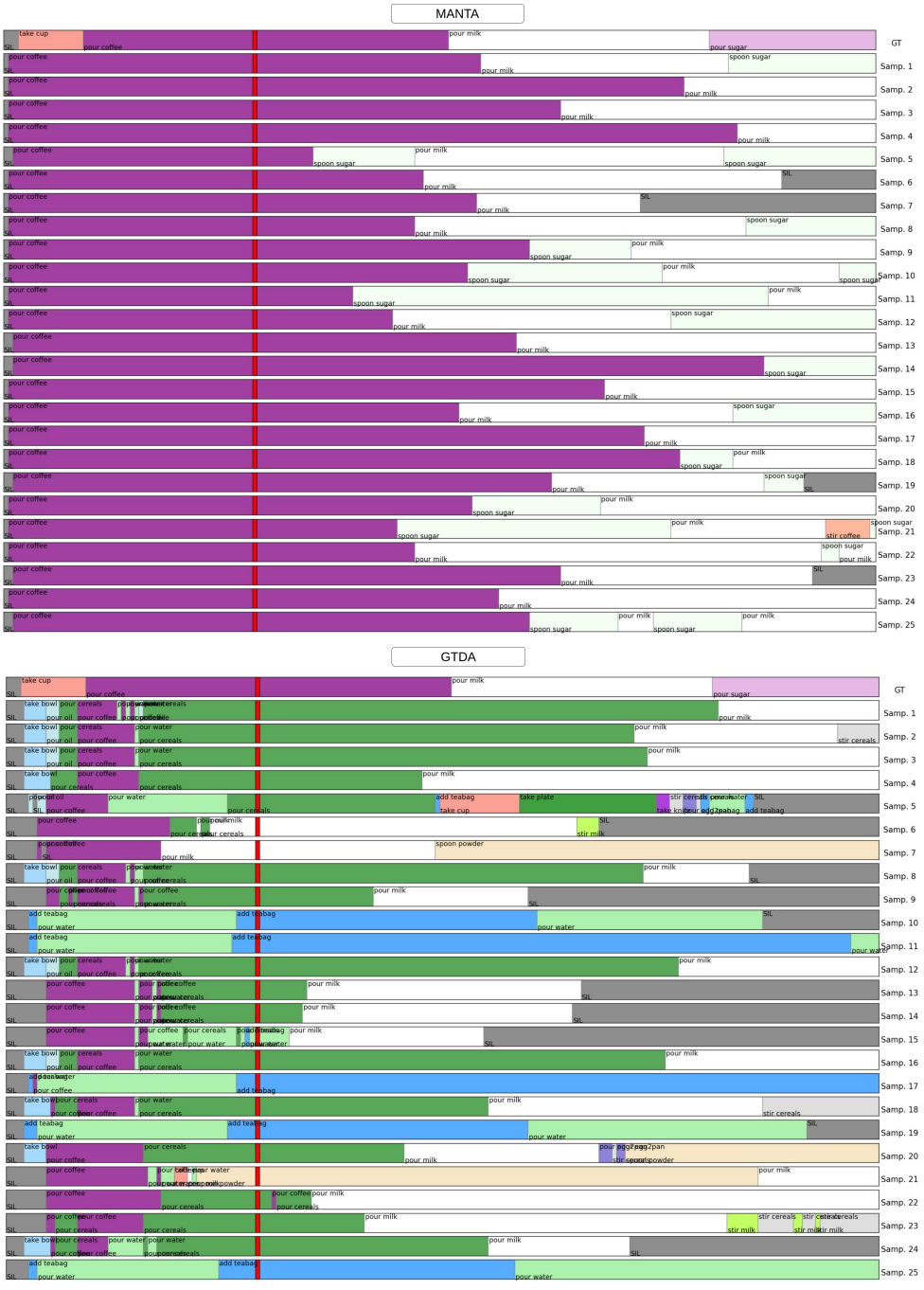

Figure 3. Qualitative comparison of MANTA *(top)* to GTDA [2] *(bottom)* on the Breakfast dataset. Best viewed zoomed in.

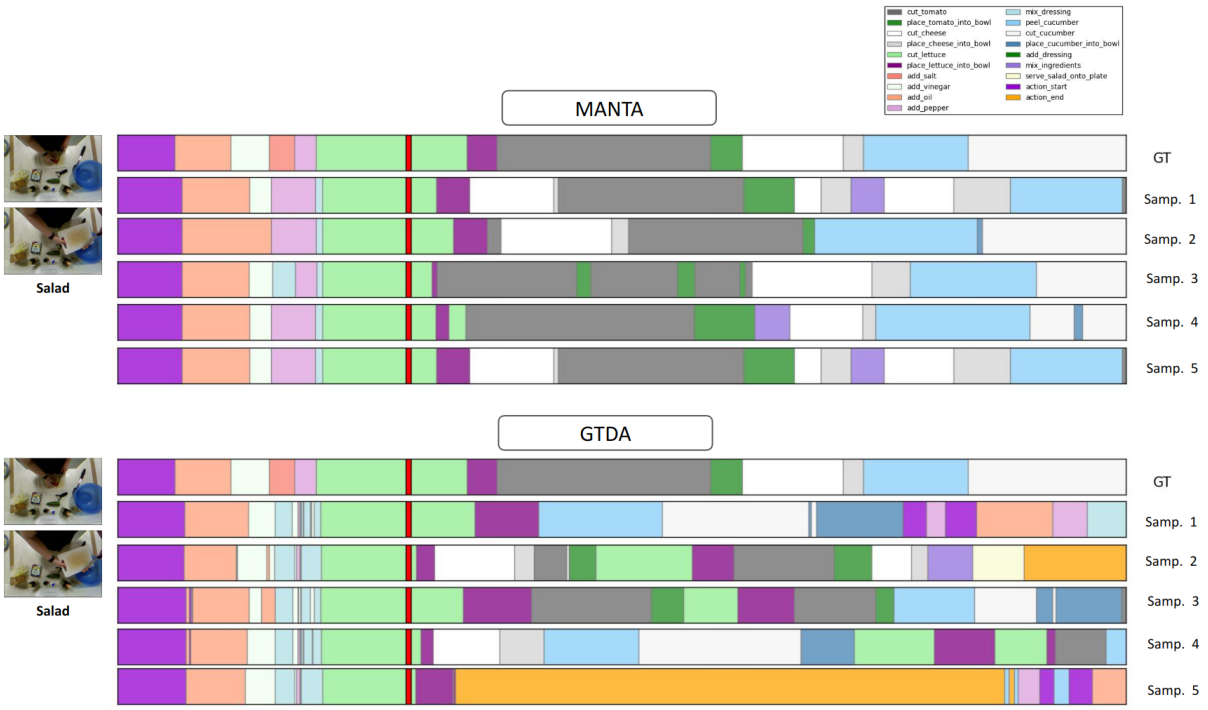

Figure 4. Qualitative comparison of MANTA *(top)* to GTDA [2] *(bottom)* on the 50Salads dataset. Best viewed zoomed in.

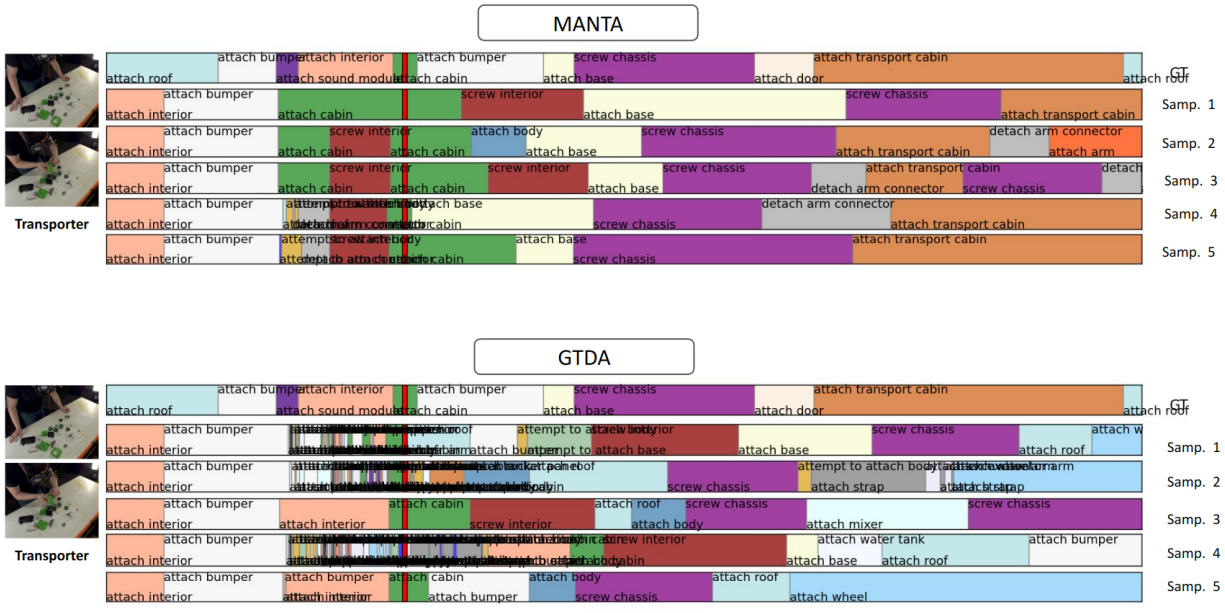

Figure 5. Qualitative comparison of MANTA *(top)* to GTDA [2] *(bottom)* on the Assembly101 dataset. Best viewed zoomed in.