Supplementary Material

A. Long-tail Temporal Action Segmentation

The long tail problem has been overlooked in temporal action segmentation. State-of-the-art methods [15, 35, 52] perform poorly and often do not predict any tail classes correctly. For example, MSTCN [15], ASFormer [52], and DiffAct [35] how zero accuracies for 5, 5, and 4 out of 48 classes on the Breakfast dataset as shown in the per-class accuracy plot in Fig. 9. Our paper is the first to address the long tail problem in temporal action segmentation.

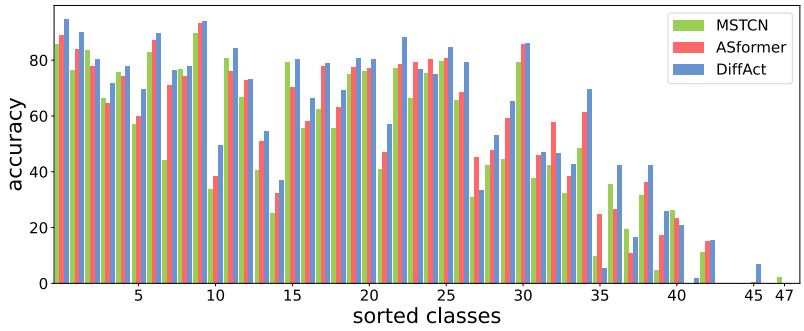


Fig. 9: Class-wise accuracy distribution on Breakfast with MSTCN, ASFormer and DiffAct. Classes are sorted by their frame counts in the training set.

B. Grouping Without Activity Labels

During training, our group-wise classification relies on grouping classes based on activity labels. However, in scenarios where the activity label is unavailable or should not be utilized, an alternative approach is forming groups through sequence clustering. Notably, clustering results often align with the underlying activity label. Some group examples are shown in Tab. 9. During testing, we use Eq. (13) to identifyied groups. Details of the group identification results can be found in Sec. E.

Table 9: Grouping examples.

Dataset	Group	Action Classes
Breakfast	*"coffee"*	*take_cup*, *pour_sugar*, *spoon_sugar*, *SIL*, *pour_coffee*, *stir_coffee*, *pour_milk*
GTEA	*"Cluster 1"*	*Pour*, *Take*, *Close*, *Put*, *Open*, *Fold*, *Background*

Clustering algorithm We cluster sequences according to the action frequency distribution, leveraging its ability to capture the action co-occurrence patterns [13]. Given a sequence i, the action frequency distribution q is defined as the normalized occurrence of frames for all the actions:

$$q_i(c) = \frac{1}{T_i} \sum_{t}^{T_i} \mathbb{1}(y_t = c), \quad c \in [1, \cdots L]. \tag{14}$$

Then, we can define the distance criterion between two sequences i, j using the Kullback-Leibler(KL)-divergence:

$$dist(i, j) = \frac{1}{2} \sum_{c} q_i(c) \log \frac{q_i(c)}{q_j(c)} + q_j(c) \log \frac{q_j(c)}{q_i(c)} \tag{15}$$

We average over the forward and backward KL-divergence to ensure symmetry in the distance measure. Based on the defined distance criterion, we apply hierarchical clustering [23] with a predefined number of groups and a tuned linkage criterion to establish the sequence-to-group mapping.

Effect of the number of groups n. We present results of using clustering on Breakfast with MSTCN to assess the impact of the number of groups. Setting the number of clusters to $n = 10$ yields the same groups as using the activity label. Further reducing n leads to the merging of several activities. For instance, setting $n = 8$ merges the activities *"cereal"-"milk"* and *"friedegg"-"scrambledegg"*. According to the results in Tab. 10, finer clustering, *i.e.* more detailed separation of the activities, contributes to better performance by reducing the false positives from activity-irrelevent classes to a larger extent.

Table 10: Varying the number of groups n for group-wise classification, with fixed $\eta = 0.5$ and $\tau = 0.5$

n	Frame acc			Seg. F1		
	Head	Tail	Hmean	Head	Tail	Hmean
3	65.9	39.7	49.6	56.3	40.3	47.0
5	66.3	41.1	50.7	**61.2**	43.4	50.8
8	66.5	41.5	50.9	60.3	44.5	51.2
10	**67.6**	**43.0**	**52.7**	60.1	**45.2**	**51.5**

C. Temporal Priors

For an action c, we defined two sets $S_{bf}[c]$ and $S_{af}[c]$ that contain actions that must precede and follow action c. These two sets are utilized to fix the temporal bounds when employing logit adjustment on action c. These two sets are exclusive and can be extracted from the training data. The algorithm for finding these two sets for a given class c is given in Algorithm 1. Examples of extracted temporal bounds for actions on Breakfast can be found in Tab. 11

Algorithm 1 Exacting Temporal Bounds for Class c

Input: Training sequence labels $\mathcal{Y}$, Class c
Initialize: $A = \text{set}()$, $B = \text{set}()$
for Y in $\mathcal{Y}$ **do** *// for each sequence*
  if $c \notin Y$ **then**
    continue
  $ls = \text{get\_seg\_label}(Y)$ *// segment-wise label*
  $ids = ls.\text{index}(c)$
  for $i \in ids$ **do**
    $A = A \cup ls[i{+}1:]$ *// update actions after c*
    $B = B \cup ls[:i]$ *// update actions before c*
$S_{bf}[c] = B - A$ *// must precede c*
$S_{af}[c] = A - B$ *// must follow c*
Return: $S_{bf}[c],\ S_{af}[c]$

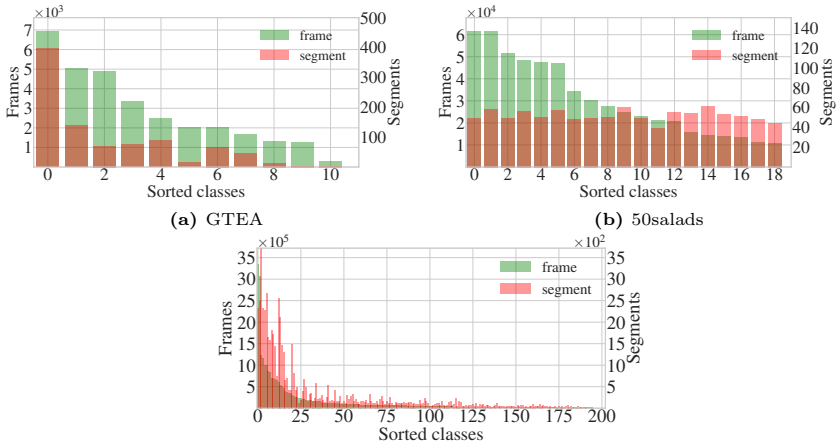
Table 11: Examples of extracted temporal bounds for actions in Breakfast dataset.

Actions	$S_{bf}[c]$	$S_{af}[c]$
pour cereals	*take bowl*	*pour milk, stir cereals*
take bowl	-	*pour milk, pour cereals, stir cereals*
add teabag	*take cup*	*pour sugar, spoon sugar, stir tea*
stir milk	*take cup*	-

D. Experimental Setting

Dataset. We conduct experiments on five benchmarks. (1) Breakfast consists of 1712 videos with ten video-level activities for making breakfast. On average, the videos are 2.3 minutes long with 48 action classes. (2) YouTube Instructional is a collected dataset that includes five instructional activities. It comprises 30 videos for each activity, totaling 46 unique action classes. (3) Assembly101 is a recently collected dataset for dissembling and assembling take-apart toys. It has a collection of 4321 videos with an average length of 7.1 minutes and 202 coarse actions. (4) GTEA contains 28 videos of seven procedural activities recorded in a single kitchen, with a total of 11 actions. (5) 50Salads is composed of 50 recorded videos of making mixed salads involving 19 actions. Despite the less imbalanced nature, we include the GTEA and 50Salads results for comprehensive evaluation, as these datasets are widely utilized within the research community. Data distribution of these datasets is illustrated in Fig. 10.

Implementation Details. The experimental configurations of base models are summarized in Tab. 12. All models are trained to reduce over-segmentation with an extra smoothing loss [15] with $\lambda = 0.15$. We follow the protocols in original papers for any details not specified here. As MSTCN and ASFormer are multi-stage models, the long-tail methods are exclusively applied at the final stage. Applying these methods to all stages results in degraded performance. We include several types of methods for comparison.



(a) GTEA **(b)** 50salads

(c) Assembly101

Fig. 10: Data distribution of Assembly101, GTEA, and 50salads.

Table 12: Base model setup

Model	Optimizer	lr	weight decay	Batch size	Epochs	Sample rate
MSTCN	Adam	5×10^{-4}	-	1	50	1
ASFormer	Adam	1×10^{-4}	1×10^{-5}	1	60	4

- **Re-weighting.** Focal [34] assigns different weights to samples based on their difficulty; CB [11] calculates the weight for each class based on its effective number of samples.
- **Logit adjustment.** LA [37] and LDAM [7] both adjust the logits based on the class prior, where the prior is estimated with the class-independent assumption. Seesaw [48] dynamically re-balances the gradients of positive and negative samples by adjusting the logits.
- **Post-hoc process.** τ-norm [24] normalizes the weights of a learned classifier to achieve a balanced classifier.
- **Ensemble.** BAGS [33] utilizes group-wise training by modulating the training for head and tail classes separately to ensures both are sufficiently trained.

Hyperparameters used in above methods are selected through grid search. Hyperparameters that yield the best overall balanced and global metrics results are selected. Our method adopts the hyperparameter η to balance the target and non-target group losses, and τ as in LA [37] to tune the head-tail trade-off. Tab. 13 gives the search space for different methods. Please refer to the references for the meaning of hyperparameters in each method.

Table 13: Search space of hyperparameters

Method	Search space
Focal [34]	$\gamma \in \{0.5, 1.0, 1.5\}$
CB [11]	$\beta \in \{0.9, 0.99, 0.999\}$
LA [37]	$\tau \in \{0.1, 0.3, 0.5, 0.7\}$
LDAM [7]	$s \in \{1, 3, 5, 10\}$, $m \in \{0.5, 1.0, 1.5\}$
Seesaw [48]	$p \in \{0.1, 0.2, 0.4\}$, $q \in \{0.5, 1.0, 1.5\}$
τ-norm [24]	$\tau \in \{0.5, 1.0, 1.5\}$
BAGS [33]	$\beta \in \{2, 4, 8\}$, $N \in \{2, 3, 4\}$
G-TLA(ours)	$\eta \in \{0.1, 0.3, 0.5, 1.0\}$

The hyperparameters used for each dataset, backbone, and method are detailed in Table 14. We omit τ-norm in the tables as the results always favor $\tau = 1.0$ for τ-norm. Our approach adopts a group-wise classification framework, where the number of groups is decided either based on activity label or by clustering: for Breakfast and Youtube, we group based on activity labels with n set to 10 and 5, respectively; for GTEA and Assembly, clustering forms groups with n set to 3 and 2, respectively; for 50salads, group-wise classification is discarded, equivalent to $n = 1$.

Table 14: Hyperparameters for different experimental settings.

Data	Model	Focal [34] γ	CB [11] β	LA [37] τ	LDAM [7] s, m	Seesaw [48] p, q	BAGS [33] β, N	G-TLA η, τ
Breakfast	MSTCN	0.5	0.9	0.5	1.0, 0.5	0.4, 0.5	8, 3	0.5, 0.5
	ASFormer	1.5	0.9	0.1	1.0, 1.5	0.1, 0.5	8, 3	0.1, 0.3
YouTube	MSTCN	1.0	0.9	0.3	1.0, 0.5	0.1, 0.5	4, 3	0.1, 0.5
	ASFormer	1.0	0.9	0.3	1.0, 1.5	0.1, 1.5	8, 3	0.1, 0.3
GTEA	MSTCN	-	0.999	0.7	-	-	-	1.0, 0.5
	ASFormer	-	0.9	0.5	-	-	-	1.0, 0.1
50salads	MSTCN	-	0.9	0.5	-	-	-	-, 0.3
	ASFormer	-	0.99	0.7	-	-	-	-, 0.3
Assembly101	MSTCN	-	0.99	0.1	-	-	-	1.0, 0.1
	ASFormer	-	0.9	0.3	-	-	-	1.0, 0.3

Table 15: Comparisons on YouTube with harmonic mean of head and tail classes over 3 runs.

Model	Frame acc			Segment F1@25		
	Head	Tail	Hmean	Head	Tail	Hmean
AsFormer	53.1	17.2	26.0	47.6	20.2	28.4
+ G-TLA(ours)	**55.4**$_{\pm 0.8}$	**24.0**$_{\pm 0.6}$	**33.5**$_{\pm 0.5}$	**47.3**$_{\pm 0.6}$	**25.3**$_{\pm 0.7}$	**33.0**$_{\pm 0.6}$
MSTCN	46.0	15.5	23.2	39.0	16.8	23.5
+ G-TLA(ours)	**48.7**$_{\pm 1.0}$	**21.8**$_{\pm 0.8}$	**30.0**$_{\pm 0.8}$	**41.7**$_{\pm 0.4}$	**20.1**$_{\pm 0.5}$	**27.1**$_{\pm 0.5}$

Table 16: Comparison on Breakfast with harmonic mean on the head and tail classes over 3 runs.

Model	Frame acc			Segment F1@25		
	Head	Tail	Hmean	Head	Tail	Hmean
AsFormer	69.7	39.8	50.7	69.9	43.9	53.9
+ G-TLA(ours)	**70.3**$_{\pm 0.1}$	**43.2**$_{\pm 0.5}$	**53.3**$_{\pm 0.5}$	**71.7**$_{\pm 0.2}$	**46.5**$_{\pm 0.1}$	**56.5**$_{\pm 0.1}$
MSTCN	65.1	37.7	47.7	53.3	38.7	44.8
+ G-TLA(ours)	**67.6**$_{\pm 0.4}$	**43.0**$_{\pm 0.6}$	**52.7**$_{\pm 0.6}$	**60.1**$_{\pm 0.8}$	**45.2**$_{\pm 0.4}$	**51.5**$_{\pm 0.6}$

Table 17: Global and balanced result summary for Breakfast.

Model	Global					Balanced			
	Edit	Acc	F1@10	F1@25	F1@50	Acc	F1@10	F1@25	F1@50
MSTCN	66.6	67.7	63.2	57.9	46.0	47.7	48.3	44.8	36.9
+ CB [11]	66.8	67.4	63.6	57.9	45.7	48.8	49.2	45.6	37.3
+ Focal [34]	67.3	68.5	63.1	57.5	45.5	46.7	48.4	44.4	35.6
+ BAGS [33]	66.3	68.5	65.1	59.8	47.5	49.4	51.1	47.4	38.5
+ τ-norm [24]	66.3	67.9	62.4	57.0	45.1	46.6	47.1	43.8	35.8
+ LA [37]	67.2	67.6	63.1	57.9	45.6	49.8	49.0	45.7	36.8
+ LDAM [7]	67.1	67.5	63.4	58.1	46.1	48.0	49.1	45.6	37.4
+ Seesaw [48]	67.4	68.6	63.1	57.8	46.2	50.1	48.8	45.3	37.2
+ G-TLA(ours)	**71.3**	**70.3**	**68.3**	**62.9**	**50.0**	**52.7**	**54.5**	**51.5**	**41.5**
ASFormer	74.5	72.4	75.5	69.9	56.1	50.7	57.1	53.9	44.6
+ CB [11]	74.9	71.9	75.6	69.7	55.8	51.6	57.9	54.9	45.6
+ Focal [34]	75.4	72.3	76.1	70.4	56.2	50.2	58.1	55.1	44.9
+ BAGS [33]	73.7	71.8	74.7	68.9	55.9	51.2	58.0	54.9	45.7
+ τ-norm [24]	73.6	72.2	74.9	69.1	55.7	51.5	57.1	54.2	45.2
+ LA [37]	74.9	72.5	75.6	69.7	56.3	51.3	58.6	54.9	46.0
+ LDAM [7]	75.3	**72.6**	76.0	70.6	**57.2**	51.7	58.2	55.1	46.6
+ Seesaw [48]	74.9	72.5	75.7	70.1	56.3	51.8	58.6	55.5	45.8
+ G-TLA(ours)	**75.5**	72.2	**76.2**	**70.9**	56.8	**53.3**	**59.2**	**56.5**	**47.5**

E. Additional Results

Main results. We provide benchmark results with standard deviation over 3 runs in Tab. 16 and Tab. 15. Each run is over 4 or 5 splits(depending on the

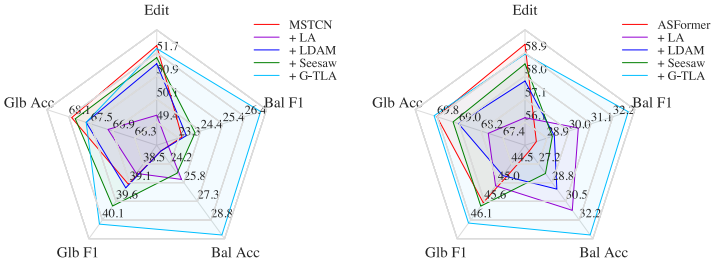
dataset). We also show the detailed results on global metrics(Acc, F1@10, F1@25, F1@50, and edit score) and balanced metrics(Acc, F1@10, F1@25, F1@50) in Tab. 17 and Tab. 18. Our method consistently outperforms others across datasets and backbones. These results affirm the effectiveness of our approach in mitigating over-segmentation while concurrently improving balanced metrics. The additional results on F1 scores at IoU thresholds of 0.10 and 0.50 further reveal the consistent trend of F1 scores at different thresholds across various methods.

Table 18: Global and balanced result summary for Youtube.

Model	Global					Balanced			
	Edit	**Acc**	**F1@10**	**F1@25**	**F1@50**	**Acc**	**F1@10**	**F1@25**	**F1@50**
MSTCN	51.9	**68.0**	44.7	39.1	23.7	23.2	27.5	23.5	15.0
+ CB [11]	50.9	66.3	44.9	38.7	24.0	25.9	27.7	23.3	15.7
+ Focal [34]	**52.1**	67.5	45.4	40.0	24.3	25.0	28.6	25.4	15.5
+ BAGS [33]	50.8	67.2	45.7	40.1	**25.4**	25.3	28.6	24.4	15.8
+ τ-norm [24]	50.0	67.3	43.8	38.0	23.6	24.5	26.7	22.9	15.4
+ LA [37]	49.6	67.0	44.4	38.8	23.2	25.5	27.2	22.8	14.4
+ LDAM [7]	51.3	67.6	45.2	39.2	23.6	23.2	28.0	23.7	13.7
+ Seesaw [48]	51.5	67.9	45.4	39.7	23.2	25.0	27.6	24.1	14.0
+ G-TLA(ours)	51.7	67.6	**45.8**	**40.2**	24.9	**30.0**	**30.9**	**27.1**	**17.2**
ASFormer	**59.8**	69.8	51.8	45.6	29.1	26.0	32.7	28.4	19.1
+ CB [11]	57.7	69.6	51.3	45.4	28.7	28.8	33.7	29.2	19.4
+ Focal [34]	59.3	69.7	52.5	**46.6**	29.8	26.1	35.2	29.8	18.2
+ BAGS [33]	57.0	69.3	51.1	45.1	29.6	29.3	34.1	30.0	20.4
+ τ-norm [24]	58.4	69.0	50.8	44.3	28.7	27.6	32.2	28.5	19.0
+ LA [37]	56.3	67.9	51.3	45.1	27.5	31.3	35.0	30.5	19.7
+ LDAM [7]	57.8	69.0	51.1	44.8	28.4	29.4	33.8	29.3	19.2
+ Seesaw [48]	58.5	69.2	52.0	45.7	29.4	28.0	33.5	29.2	18.1
+ G-TLA(ours)	58.9	**69.9**	**52.8**	46.2	**29.9**	**33.5**	**38.8**	**33.0**	**22.5**

Extra plots for YouTube. We further visualize the global and per class results using radar charts on YouTube dataset in Fig. 11, specifically comparing the performance of logit adjustment methods. The plot demonstrates the superior performance of our method, indicated by the largest enclosed area. Our method excels in segment-wise performance, including edit score and global & balanced F1 score.

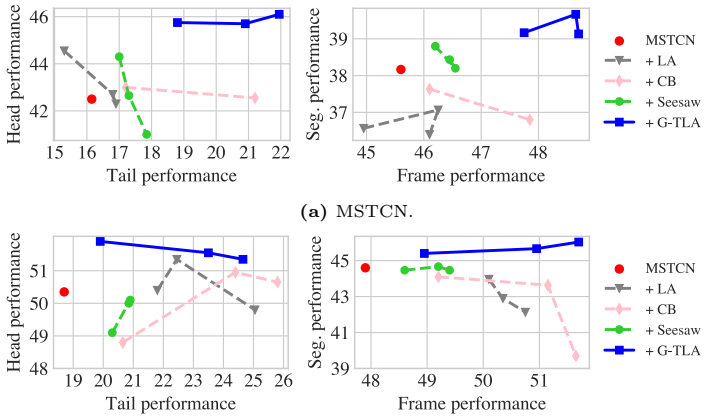
Long-tail methods for temporal action segmentation exhibit two primary trade-offs: the head-tail trade-off, negatively impacting the head when improving the tail, and the frame-segment trade-off, wherein enhancing tail might adversely affects segment-wise performance. These trade-offs is directly influenced by the hyperparameters. We show these two trade-offs for various methods across backbones in Fig. 12. We fix $\eta = 0.1$ and change $\tau \in \{0.1, 0.3, 0.5\}$ for the trend plotting of our method. For CB and LA, $\beta \in \{0.9, 0.99, 0.999\}$ and $\tau \in \{0.3, 0.5, 0.7\}$.



(a) MSTCN. **(b)** AsFormer.

Fig. 11: Radar charts of logit adjustment methods, measuring the performance along balanced and global metrics on YouTube with MSTCN and AsFormer.

For Seesaw, q is fixed as 0.5, and $p \in \{0.1, 0.2, 0.3\}$. Notably, the curve of our method consistently outperforms others , emphasizing its effectiveness in balancing the learning between head and tail, as well as mitigating over-segmentation.



(a) MSTCN.

(b) AsFormer.

Fig. 12: Head-Tail & Frame-Segment trade-offs on YouTube with MSTCN and AsFormer.

Ablation study. We present additional ablation studies on YouTube, highlighting the contributions of each component of our G-TLA in Tab. 19. Compared to groupwise classification, temporal logit adjustment yields greater improvement, emphasizing the importance of incorporating temporal priors for YouTube.

The effect of the hyperparameter η on YouTube with MSTCN and AsFormer is shown in Tab. 20 and Tab. 21. Small η reduces suppression of tail classes, but if too small, it harms group identification during inference. Conversely, a large η over-emphasizes the 'others' class, harming tail performance.

Tab. 22 and Tab. 23 present the effect of the hyperparameter τ on YouTube. A smaller τ represents minimal adjustment, resulting in less improvement for tail classes. Conversely, a large value of τ biases towards tail classes and introduces more false positives, causing more over-segmentation.

Table 19: Ablate group classification(GP), logit adjustment(LA), temporal factor(TF) on YouTube.

GP	LA	TF	MSTCN						ASFormer					
			Frame acc			Seg. F1			Frame acc			Seg. F1		
			Head	Tail	Hmean	Head	Tail	Hmean	Head	Tail	Hmean	Head	Tail	Hmean
✗	✗	✗	46.0	15.5	23.2	39.0	16.8	23.5	53.1	17.2	26.0	**47.6**	20.2	28.4
✗	✓	✗	46.0	17.6	25.5	39.4	16.0	22.8	53.9	22.1	31.3	46.9	22.5	30.5
✓	✗	✗	46.5	16.1	23.9	**42.5**	16.8	24.0	53.1	17.6	26.4	46.9	21.8	29.7
✓	✓	✗	**48.9**	19.9	28.3	42.3	17.7	25.0	55.3	20.6	30.0	47.1	21.8	29.8
✓	✓	✓	48.7	**21.8**	**30.0**	41.7	**20.1**	**27.1**	**55.4**	**24.0**	**33.5**	47.3	**25.3**	**33.0**

Group identification results. Group identification impacts the final performance. Group identification accuracy is shown in Tab. 24. The baseline selects predicted groups based on summed output probabilities within each group. Our groupwise classification improves activity identification, reducing false positives and enhancing both frame and segment-wise performance. For datasets without activity label, we use clustering results as ground truth. Our method improves group accuracy from 62.0%(baseline) to 83.0% on Assembly101 using MSTCN.

Results on other datasets In Tab. 25, Tab. 26, and Tab. 27, we present our model's performance on other three datasets for completeness. GTEA [17] and 50Salads [46] have smaller vocabulary sizes and are less imbalanced. Assembly101 [41] is long-tailed but less explored. We do not emphasize Assembly101 as the head classes are also not well-learned. Enhancing the tail classes is less meaningful when the head classes still perform poorly. We include several methods from each type to show the comparison. Our method demonstrates competitive results, further validating its effectiveness.

Table 20: Varying η for group-wise classification, with fixed number of groups $n = 5$ and $\tau = 0.5$ on YouTube with MSTCN.

η	Frame acc			Seg. F1		
	Head	Tail	Hmean	Head	Tail	Hmean
0.1	**48.7**	**21.8**	**30.0**	41.7	**20.1**	**27.1**
0.3	48.4	18.5	26.7	42.2	16.4	23.6
0.5	48.4	18.3	26.5	**43.0**	16.9	24.2

Table 21: Varying η for group-wise classification, with fixed number of groups $n = 5$ and $\tau = 0.3$ on YouTube with AsFormer.

τ	Frame acc			Seg. F1		
	Head	Tail	Hmean	Head	Tail	Hmean
0.1	55.4	**24.0**	**33.5**	47.3	**25.3**	**33.0**
0.3	**55.6**	21.6	31.1	48.4	21.8	30.1
0.5	55.4	19.3	30.0	**48.6**	21.5	29.8

Table 22: Varying τ for temporal logit adjustment, with fixed number of groups $n = 5$ and $\eta = 0.1$ on Youtube with MSTCN.

η	Frame acc			Seg. F1		
	Head	Tail	Hmean	Head	Tail	Hmean
0.1	49.4	17.1	25.4	41.1	18.2	25.3
0.3	48.8	18.0	26.0	41.7	18.2	25.7
0.5	48.7	21.8	30.0	41.7	**20.1**	**27.1**
0.7	**48.9**	**22.5**	**30.8**	**42.2**	17.9	25.4

Table 23: Varying τ for temporal logit adjustment, with fixed number of groups $n = 5$ and $\eta = 0.1$ on Youtube with AsFormer

τ	Frame acc			Seg. F1		
	Head	Tail	Hmean	Head	Tail	Hmean
0.1	54.3	18.9	27.9	**49.5**	21.1	29.7
0.3	**55.4**	**24.0**	**33.5**	47.3	**25.3**	**33.0**
0.5	**55.4**	23.4	32.9	46.7	23.6	30.6

Table 24: Accuracy of group identification

Method	Breakfast		Youtube	
	MSTCN	ASFormer	MSTCN	ASFormer
Baseline	87.2	89.1	89.3	93.1
G-TLA	90.1	90.2	93.4	94.2

Table 25: Additional results on 50salads.

Model	Frame acc			Segment F1@25			Global		
	Head	Tail	Hmean	Head	Tail	Hmean	Edit	F1@25	Acc
AsFormer	90.6	77.4	83.5	87.5	80.3	83.8	79.0	82.3	85.2
+ CB [11]	90.9	78.1	84.0	88.4	81.4	84.8	78.7	83.2	85.8
+ τ-norm [24]	90.4	77.5	83.5	87.7	80.2	83.8	78.9	82.2	85.2
+ LA [37]	90.2	78.3	83.8	**89.8**	82.1	85.7	79.8	84.5	85.4
+ G-TLA(ours)	**90.8**	**79.7**	**84.9**	89.4	**83.1**	**86.1**	**80.7**	**84.6**	**86.3**
MSTCN	87.7	70.0	77.9	85.7	72.1	78.3	71.4	75.9	81.1
+ CB [11]	88.4	69.3	77.7	85.3	72.0	78.1	71.1	75.5	81.0
+ τ-norm [24]	87.6	70.3	78.0	85.1	71.6	77.8	70.8	75.3	81.1
+ LA [37]	87.5	69.6	77.5	86.0	71.0	77.8	70.4	75.2	80.8
+ G-TLA(ours)	**89.0**	**71.7**	**79.4**	**86.8**	**73.8**	**79.8**	**72.0**	**77.3**	**81.9**

Table 26: Additional results on GTEA.

Model	Frame acc			Segment F1@25			Global		
	Head	Tail	Hmean	Head	Tail	Hmean	Edit	F1@25	Acc
AsFormer	**80.6**	81.7	81.2	72.5	85.4	78.4	**88.4**	89.4	81.1
+ CB [11]	79.5	84.0	81.6	71.4	88.3	78.9	86.7	89.0	80.8
+ τ-norm [24]	80.5	82.1	81.3	**72.8**	85.2	78.5	88.3	89.5	81.1
+ LA [37]	79.5	82.7	81.1	70.9	88.4	78.7	87.5	88.9	80.5
+ G-TLA(ours)	80.2	**84.5**	**82.3**	72.0	**90.4**	**80.2**	87.9	**89.6**	**81.2**
MSTCN	**77.6**	80.3	78.9	69.4	86.6	77.0	84.8	87.2	78.0
+ CB [11]	76.2	82.6	79.3	68.8	90.1	78.0	85.3	87.7	78.5
+ τ-norm [24]	77.5	80.6	79.0	69.2	86.8	77.0	84.5	87.2	78.0
+ LA [37]	77.0	83.0	79.8	**69.8**	86.3	77.2	85.4	87.2	78.5
+ G-TLA(ours)	77.5	**83.7**	**80.5**	69.5	**90.3**	**78.5**	**85.8**	**87.9**	**78.6**

Table 27: Additional results on Assembly101.

Model	Frame acc			Segment F1@25			Global		
	Head	Tail	Hmean	Head	Tail	Hmean	Edit	F1@25	Acc
AsFormer	35.2	5.7	9.8	29.0	4.8	8.2	**31.8**	**30.4**	41.1
+ CB [11]	35.4	5.9	10.1	26.5	5.2	8.7	30.6	28.2	41.0
+ τ-norm [24]	32.2	4.9	8.5	21.8	3.2	5.6	24.3	22.7	38.5
+ LA [37]	36.1	5.9	10.1	27.5	5.7	9.4	30.2	28.5	**41.4**
+ G-TLA(ours)	**36.8**	**9.2**	**14.7**	**30.7**	**8.3**	**13.1**	30.7	29.8	41.0
MSTCN	33.9	4.7	8.2	26.3	3.9	6.8	30.1	27.2	**39.8**
+ CB [11]	32.5	7.3	11.9	27.1	4.5	7.7	25.1	22.4	37.9
+ τ-norm [24]	34.0	4.3	7.6	25.9	4.2	7.2	30.5	27.4	39.6
+ LA [37]	34.1	7.4	12.1	27.3	5.6	9.3	30.0	26.2	39.5
+ G-TLA(ours)	**34.9**	**8.0**	**13.0**	**30.2**	**5.8**	**9.7**	**30.5**	**28.5**	39.2