A. Terminology, Notation, and Explanation

In the table below, we list all the terminologies used in this work, their notations, and concise explanations.

Terminology	Notation	Explanation	Value in Fig. 5
Sequence len	N	An integer.	10
# workers	W	An integer.	7
Worker index	Wk_idx	An integer. The worker whose $Wk\_idx = i$ is denoted as $\mathtt{Wk}_i$.	$0, 1, \cdots, 6$
Interest set	$\mathcal{g}$	A list of integers. It is the CQS base set starting with 0 and 1, and has all-pairs property.	$[0, 1, 3]$
Quorum size	m, m_{TGi} m_{Tki}	An integer. m_{TGi} is # token groups for Wk_i after self-distillation. $m_{TGi} \leq m$ where m = len($\mathcal{g}$) m_{Tki} is # tokens sent to Wk_i or len(subsequence). m_{Tki} is calculated from m_{TGi} based on ***TG-Tk map***.	$m = m_{TG} = 3$ $m_{Tk0} = 3$ $m_{Tk1/2/6} = 4$ $m_{Tk3/4/5} = 5$
Token index	Tk_idx	An integer. The token whose $Tk\_idx = i$ is denoted as Tk_i.	$0, 1, \cdots, 9$
Token group index	TG_idx	An integer. The token group whose $TG\_idx = i$ is TG_i.	$0, 1, \cdots, 6$
Map ratio	m_r	An integer or a float number. $m\_r = \frac{N}{W}$. Let $N = kW + r$, where $k, r \in \mathbb{Z}^+$. When r = 0, we have m_r = k. Thus, each token group (TG_i) contains k or m_r tokens. When $r \neq 0$, m_r is a float number. The first $W - r$ token groups contain k or $\lfloor m\_r \rfloor$ tokens. While each of the rest r token groups contain $k + 1$ or $\lceil m\_r \rceil$ tokens.	1.43
Material list	$\mathtt{MtrlL}_i$	A list of integers. Elements are Tk_idx NOT TG_idx. $\mathtt{MtrlL}_i$ is the list of token indices for Wk_i. Scheduler extracts the corresponding tokens (subsequence) and sends them to each worker. len($\mathtt{MtrlL}_i$) $= m_{Tki}$	e.g. $MtrlL_1 = [1, 2, 4, 5]$
Ban list	$\mathtt{BL}_i$	A list of integer pairs. Elements are Tk_idx pairs. It lists all local pairs $\mathtt{Wk}_i$ is not responsible for, hence need to be masked. Re-indexed from 0.	e.g. $BL_0 = [(1, 1), (2, 2)]$
Task list	$\mathtt{TL}_i$	A list of integer pairs. Elements are Tk_idx pairs. It lists all the pairs $\mathtt{Wk}_i$ is responsible for. Unnecessary to construct in this work, provided here for the validation purpose and future study.	e.g. $TL_0 =$ $[(0, 1), (0, 3), (1, 3),$ $(0, 0),$ $(1, 0), (3, 0), (3, 1)]$

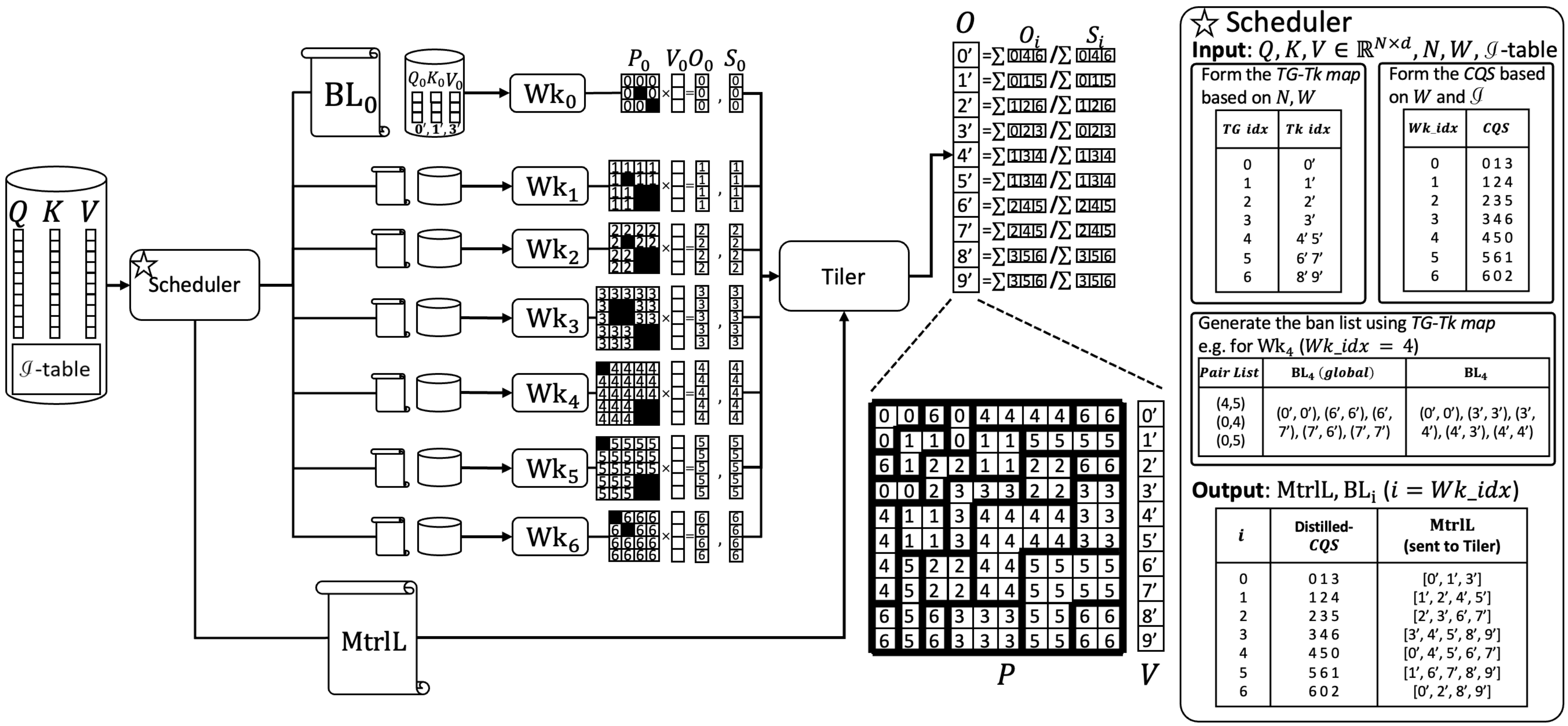
B. $\mathscr{I}$-Table for W from 3 to 111

Please see Section E for the definition of the asymptotic responsibility ratio (asym-$\mathscr{R}_R$).

W	m	Interest set ($\mathscr{I}$)	asym-$\mathscr{R}_R$
3	2	0 1	2.9675
4	3	0 1 2	3.1963
5	3	0 1 2	4.9318
6	3	0 1 3	5.1161
7	3	0 1 3	6.9095
8	4	0 1 2 4	7.0752
9	4	0 1 2 4	8.8701
10	4	0 1 2 5	9.0285
11	4	0 1 2 5	10.8493
12	4	0 1 3 7	11.0269
13	4	0 1 3 9	12.8311
14	5	0 1 5 8 13	13.0015
15	5	0 1 3 5 9	14.7823
16	5	0 1 9 11 13	15.0061
17	5	0 1 2 7 15	16.7675
18	5	0 1 4 7 9	16.8981
19	5	0 1 2 6 9	18.7381
20	6	0 1 4 7 9 10	18.8873
21	5	0 1 4 14 16	20.7349
22	6	0 1 3 7 12 20	20.9516
23	6	0 1 2 3 15 19	22.6503
24	6	0 1 11 13 17 20	22.9354
25	6	0 1 4 16 18 24	24.6391
26	6	0 1 3 6 10 14	24.8642
27	6	0 1 11 14 16 20	26.7405
28	6	0 1 4 15 20 22	26.9277
29	7	0 1 6 15 16 18 22	28.7212
30	7	0 1 4 9 11 18 24	28.9314
31	6	0 1 3 8 12 18	30.5982
32	7	0 1 7 13 23 28 31	30.8992
33	7	0 1 4 10 16 27 29	32.5885
34	7	0 1 17 21 25 29 32	32.8892
35	7	0 1 2 4 17 25 30	34.5174
36	7	0 1 3 14 20 29 32	34.898
37	7	0 1 6 8 10 23 26	36.569
38	8	0 1 11 12 30 33 35 37	36.8291
39	7	0 1 16 20 22 27 30	38.5922
40	8	0 1 6 20 23 24 33 35	38.8007
41	8	0 1 7 8 21 23 33 38	40.5428
42	8	0 1 13 22 23 26 35 37	40.8042
43	8	0 1 4 9 15 17 25 37	42.467
44	8	0 1 6 9 11 30 33 37	42.8955
45	8	0 1 5 7 9 20 23 33	44.4792
46	8	0 1 5 13 19 20 22 36	44.6193
47	8	0 1 3 6 18 19 39 43	46.3642
48	8	0 1 7 11 15 24 27 29	46.606
49	8	0 1 8 10 28 34 45 47	48.3805
50	8	0 1 17 26 36 39 44 46	48.7739
51	8	0 1 7 23 32 34 44 47	50.4064
52	9	0 1 3 8 12 23 26 36 47	50.7675
53	9	0 1 2 3 4 7 21 29 44	52.3395
54	9	0 1 2 3 4 9 15 21 31	52.6921
55	9	0 1 2 3 4 6 19 26 47	54.2648
56	9	0 1 2 3 4 11 16 33 39	54.7866
57	8	0 1 3 13 32 36 43 52	56.3451
58	9	0 1 2 3 7 21 33 37 50	56.8268
59	9	0 1 2 3 6 13 21 35 44	58.2902
60	9	0 1 2 4 9 15 25 30 42	58.6829
61	9	0 1 2 3 7 15 25 36 45	60.286
62	9	0 1 2 4 10 32 39 46 51	60.8086
63	9	0 1 2 6 8 20 38 41 54	62.2385
64	9	0 1 2 5 14 16 34 42 59	62.7657
65	9	0 1 2 6 10 28 35 51 54	64.2248
66	10	0 1 2 3 4 5 13 19 39 46	64.743
67	10	0 1 2 3 4 5 12 20 26 39	66.0142
68	10	0 1 2 3 4 10 16 21 38 45	66.6928
69	10	0 1 2 3 4 10 17 22 33 45	67.9912
70	10	0 1 2 3 4 9 20 35 49 62	68.7728
71	10	0 1 2 3 4 10 18 23 34 46	69.9969
72	10	0 1 2 3 6 11 18 31 37 51	70.6409
73	9	0 1 3 7 15 31 36 54 63	72.235
74	10	0 1 2 3 7 28 30 43 57 65	72.7827
75	10	0 1 2 5 8 18 30 32 41 56	74.0946
76	10	0 1 2 6 9 25 35 46 58 63	74.7506
77	10	0 1 2 4 10 15 37 49 56 61	76.0507
78	10	0 1 2 7 13 16 33 51 55 70	76.7324
79	10	0 1 2 6 13 28 31 47 48 71	78.1041
80	11	0 1 2 3 4 5 10 23 40 56 71	78.7409
81	11	0 1 2 3 4 5 12 20 26 39 53	79.8047
82	11	0 1 2 3 4 5 12 20 26 40 53	80.5564
83	11	0 1 2 3 4 5 12 21 27 40 54	81.8105
84	11	0 1 2 3 4 7 18 26 46 54 75	82.7315
85	11	0 1 2 3 4 9 13 25 40 54 68	83.9685
86	11	0 1 2 3 4 11 17 24 29 48 54	84.5512
87	11	0 1 2 3 4 10 42 54 62 67 73	85.8207
88	11	0 1 2 3 5 11 24 29 36 43 73	86.434
89	11	0 1 2 3 5 12 18 43 57 65 71	87.8801
90	11	0 1 2 3 6 33 46 54 67 74 81	88.7638
91	10	0 1 3 9 27 49 56 61 77 81	90.0313
92	11	0 1 2 4 40 50 51 59 64 71 77	90.4992
93	11	0 1 2 5 14 20 24 31 52 60 68	91.9202
94	12	0 1 2 3 4 5 6 14 23 30 46 61	92.4736
95	11	0 1 2 5 8 17 28 39 53 63 82	93.9306
96	12	0 1 2 3 4 5 8 21 30 53 62 86	94.6809
97	12	0 1 2 3 4 5 9 17 33 43 54 79	95.7411
98	12	0 1 2 3 4 5 11 27 40 54 69 81	96.7292
99	12	0 1 2 3 4 5 12 21 27 34 48 62	97.5321
100	12	0 1 2 3 4 5 13 20 28 34 56 63	98.4098
101	12	0 1 2 3 4 5 12 49 63 72 78 85	99.606
102	12	0 1 2 3 4 6 13 28 34 42 50 85	100.3441
103	12	0 1 2 3 4 7 38 53 62 77 85 93	101.6537
104	12	0 1 2 3 4 9 19 32 46 57 72 84	102.6145
105	12	0 1 2 3 4 10 15 36 39 61 66 89	103.7333
106	12	0 1 2 3 5 48 53 69 76 82 89 97	104.6367
107	12	0 1 2 3 5 20 27 35 42 48 58 98	105.5904
108	12	0 1 2 3 7 12 20 34 41 49 57 85	106.4
109	12	0 1 2 3 7 15 39 49 58 83 89 94	107.788
110	12	0 1 2 6 17 25 39 43 46 52 80 100	108.2861
111	12	0 1 2 5 12 27 36 38 44 52 65 93	109.7262

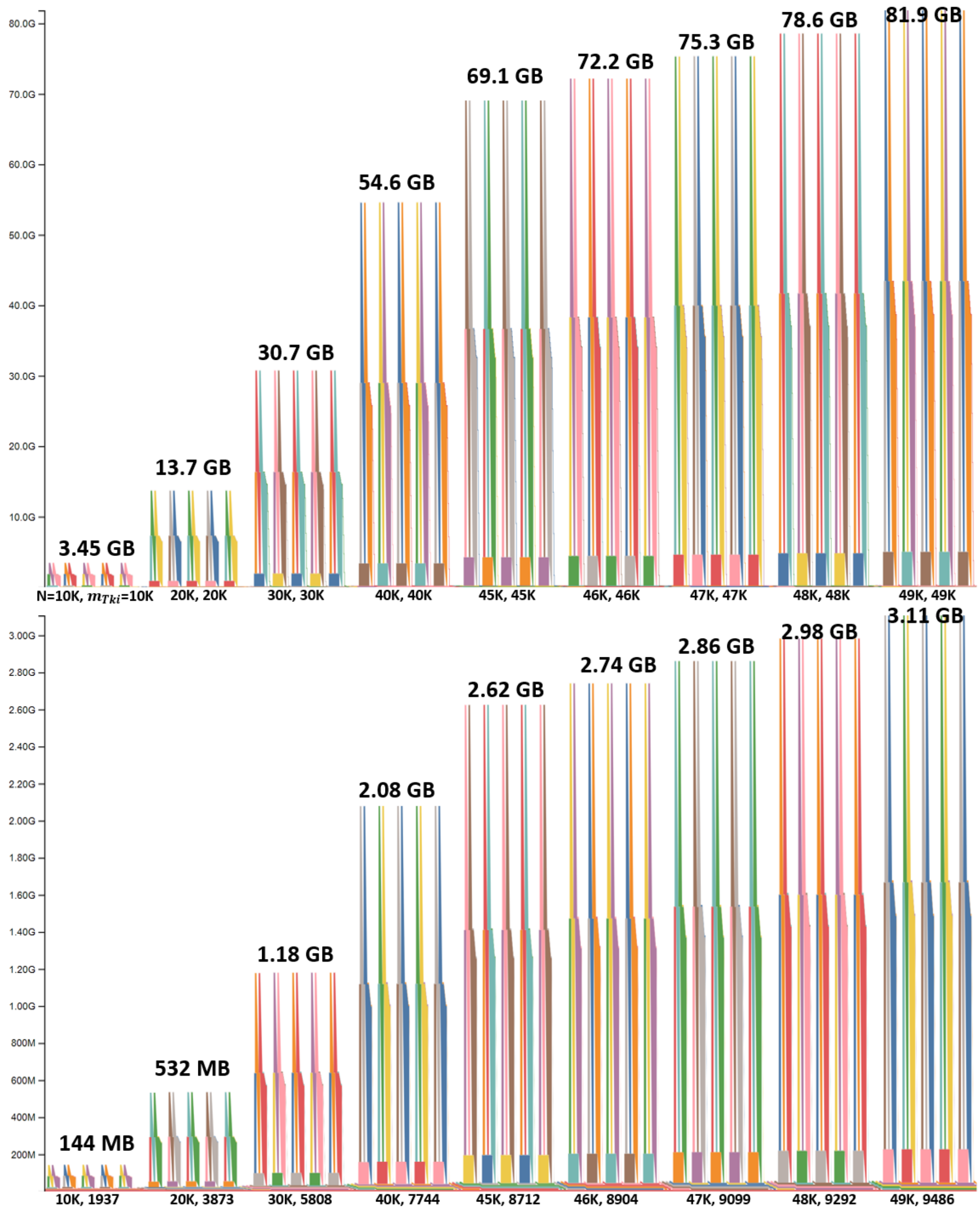
C. Workflow Example of *CQS-Attention* with Wk_idx as the Discriminative Feature (Fig. 5)

Workflow of *CQS-Attention* for $N = 10, d = 1, W = 7, \mathcal{I} = [0, 1, 3]$. Numbers in the workflow are Wk_idx, and those with prime symbol are Tk_idx. The cylinder represents data such as Q, K, V of the (sub)sequence, and $\mathcal{I}$-table. The scroll represents a ban list or material lists. Local computations include P_i, S_i, O_i but P_i need not to be sent to Tiler. Dark cells of P_i identified by the ban list ($\mathtt{BL}_i$) are zeroed. P, V are given here only for the validation purpose.



D. Active Memory Timeline of Attention Computation on NVIDIA A100 GPU

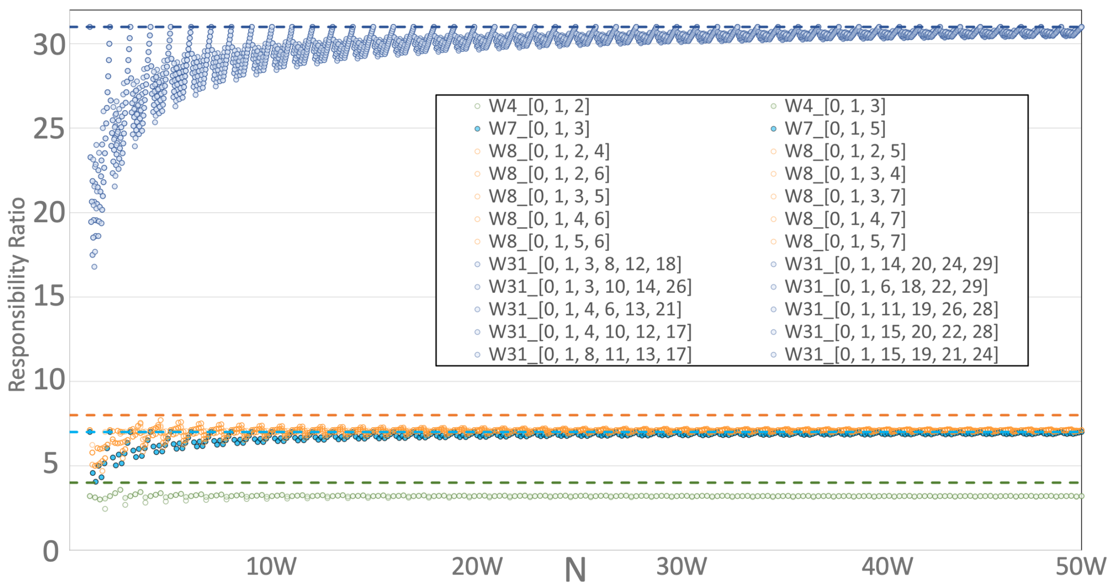
A PyTorch documentation on CUDA memory usage can be found at here. The visualizer of the memory timeline snapshot is available at pytorch.org/memory_viz. All the snapshot files of the experiment in Section IV-B can be found on our GitHub. We selectively show the $W = 1$ (upper) and $W = 31$ (lower) scenarios below. Computation is repeated 5 times.



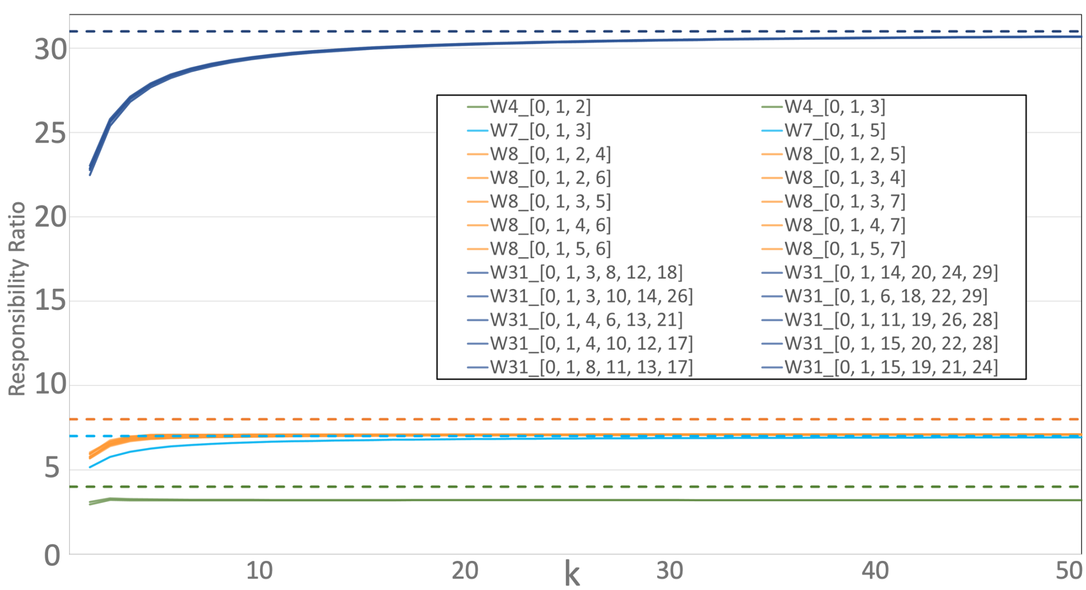
E. Comparing the Responsibility Ratio among All $\mathcal{I}$ for $W = 4, 7, 8, 31$

The workload is quadratic to the (sub)sequence length. In this section, we use a finer-grained responsibility ratio to demonstrate the difference among $\mathcal{I}$ and we adopt $\{4, 7, 8, 31\}$ as options of W to simplify the discussion. Specifically, we define the responsibility of a worker as the number of assigned cells in $P \in \mathbb{R}^{N \times N}$, and the responsibility ratio ($\mathcal{R}_R$) as N^2 over the responsibility (similar to speedup).

In the figure below, we plot $\mathcal{R}_R$ with sequence length (N) ranging from 0 to $50 \times W$, for all interest sets ($\mathcal{I}$) of $W = 4, 7, 8, 31$. Dashed lines are the corresponding optimal $\mathcal{R}_R$ (i.e. $W\times$) for each W value. As a rule of thumb, when $N < 20W$, the $\mathcal{R}_R$ difference is nontrivial. Nevertheless, as N increases, the difference becomes ignorable and $\mathcal{R}_R$ is asymptotic to a limit.



To better justify this observation, we plot the mean value of each interval, whose length is W, below. We carefully use mean value of the 41st interval ($[40W, 41W)$) to approximate this limit, dubbed asymptotic responsibility rato (asym-$\mathcal{R}_R$).



F. CQS module for General Matrix Multiplication

We put the explanation in the context of general matrix multiplication between two long sequences, which contains N_1 and N_2 tokens correspondingly. The matrix multiplication is $C = A \times B \in \mathbb{R}^{N_1 \times N_2}$, where $A \in \mathbb{R}^{N_1 \times d}$ and $B \in \mathbb{R}^{d \times N_2}$. The task is to partition and parallelize the computation of C. Identical to *CQS-Attention*, we adopt the fork-join model, and follow the Scheduler-Workers-Tiler workflow. We need to modify Scheduler in the following three aspects.

The first one is *TG-Tk map* construction. We divide each sequence to W_1 and W_2 token groups (`TG`), and $W_1 + W_2 = W$. We only look at the strictly upper part of this $W \times W$ matrix. The target $N_1 \times N_2$ matrix is equivalent to a $W_1 \times W_2$ area in this triangular part. These pairs form the task region. Left to and below the task region are `TG` pairs within each sequence.

The second modification is on the self-distillation algorithm. We not only remove `TG` pairs that the worker is not responsible for, but also those out of the task region.

Token index retrieval in the Scheduler is unnecessary hence discarded. Taking the advantage of sequential grouping and mutual exclusion, each block of token pairs is computed as a whole `TG` pair free of overlapping with any other blocks.

The duty of the Tiler is simplified: naively combining all blocks together.

For example, given $W = 7$, we divide two long sequences to 3 and 4 token groups, indexed 0-2 and 3-6. The task region, a 3×4 matrix, is highlighted below. Dashed blocks out of the task region are unnecessary. Since token index retrieval is skipped, we adopt material list, task list, and ban list to describe `TG`-pair instances. Specifically, all three `TG` pairs are banned for $\texttt{Wk}_3$, thus its material list is empty and $\texttt{Wk}_3$ is idle. The rest 6 workers compute two pairs each but stores 3 `TG`.

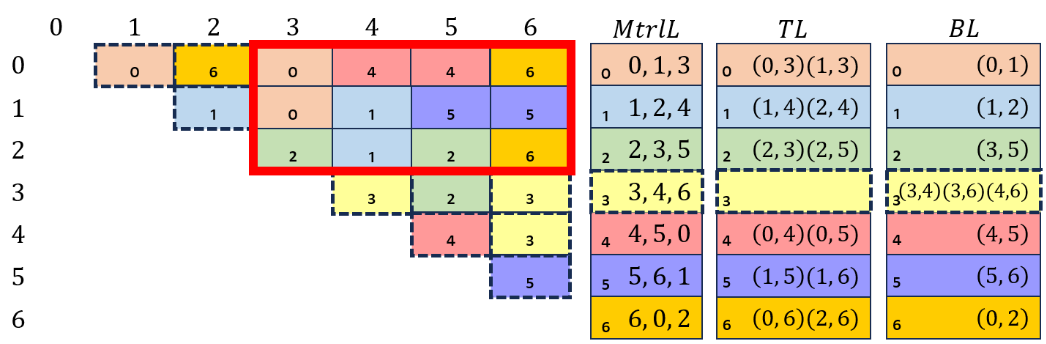


Figure 1. Matrix multiplication with 7 workers. $\texttt{Wk}_3$'s material list is empty, but filled in the figure to keep a consistent cyclic pattern.