A TEACHER COMPARISON

Method	Iteration	WMT' 16		WMT' 16	
		En-De	De-En	En-Ro	Ro-En
CMLM(Ghazvininejad et al., 2019)	10	27.03	30.53	33.08	33.31
CMLM	10	26.91	31.15	33.07	33.56
CMLMC(Huang et al., 2021)	10	28.37	31.43	34.57	34.13
CMLMC	10	27.33	30.98	33.64	33.79
Imputer(Saharia et al., 2020)	8	28.2	31.8	34.4	34.1
Imputer†	8	28.5	31.3	-	-

Table 6: Comparison of our teachers with the numbers reported in the original papers.

B EMA MOMENTUM AFFECT

We showcase the importance of the slow moving average in Figure 7. As we increase the momentum the training becomes more stable and leads to a better validation set BLEU score.

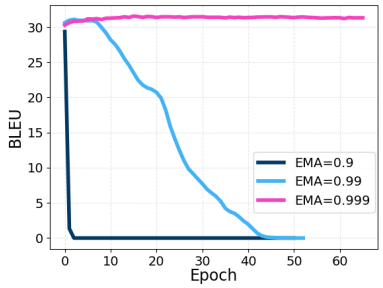

Figure 7: Validation BLEU curve on WMT'16 En-Ro for various EMA momenta with length beam=1.

C HYPER-PARAMETERS FOR DISTILLATION

We use the same hyper-parameters for all the datasets.

Hyper-parameter	CMLM/CMLMC	Imputer
Learning rate (η)	1e-3	1e-3
Adam β	(0.9, 0.98)	(0.9, 0.98)
Warmup updates	0	0
Max-tokens/GPU	8192	2048
EMA momentum (μ)	0.9992	N/A
Hidden state loss factor(λ)	0.7	0.0
Length loss factor	0.1	N/A
Mask policy	Uniform	Block
Temperature	0.5	N/A

Table 7: Summary of hyper-parameters.

D IMPUTER DETAILS

As mentioned in the main body, there is no official implementation of Imputer available online. Here, we explain the differences between our implementation and the original paper. Imputer proposes a pre-training phase where the model is optimized merely with the CTC objective. We find it unnecessary as the model reaches a better or competitive performance without it. Imputer leverages a unified decoder rather than an encoder-decoder architecture incorporated here. For Imputer training,

computing the alignment with the highest probability is necessary. This increases the training cost and Saharia et al. (2020) proposes either a pre-processing stage or using a stale copy of the active model to manage the extra computation. We compute the best alignment on the fly as it is still computationally feasible. Similar to Imputer inference, extra care is taken to make sure consecutive tokens are not unmasked in the same step. Instead of a Bernoulli masking policy during training, we used a block masking policy.

For the distillation, we observed that using the Imputer objective with hard tokens works better than KL-divergence. Imputer mainly benefits from two iterative steps and the gains are not as significant after that. Therefore, there is no incentive to use EMA.

E Ablation on Hidden State Loss Coefficient

The importance of the hidden state loss is shown in Section 4.5.1 of the main body. we conduct an ablation study in this section to find the optimal value of λ that controls the contribution of the hidden state loss.

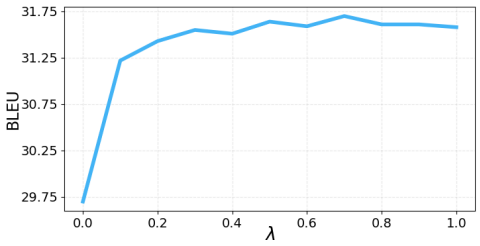

Figure 8: Best validation BLEU on WMT'16 En-Ro for CMLM with various hidden state loss coefficient (λ)

F Computational Cost

During the distillation we have to run teacher for two steps which adds extra computation. More specifically on a machine with 4 Nvidia-V100 32GB GPUs, De-En training takes approximately 11 minutes per epoch compared to 27 minutes for distillation and on En-Ro dataset training and distillation take 2 and 8 minutes per epoch, respectively. However, the number of epochs for distillation is significantly less than teacher training. Precisely, teacher training takes 250 and 200 on De-En and En-Ro datasets respectively while distillations takes 10 epochs for De-En and 30 epochs for En-Ro. Figure 9 compares the overall time for training and distillation on De-En and En-Ro datasets and it shows that the distillation time is one order of magnitude smaller than training time.

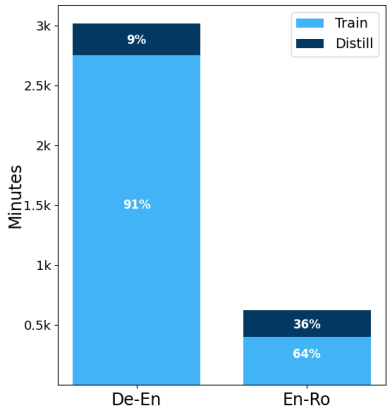

Figure 9: Comparison between teacher training and distillation time.

Note that teacher is being run in the evaluation mode, thus the activations maps are not kept in the memory. Therefore, the teacher can be run with a larger batch-size which further reduces the computational costs. We leave this as future works as it adds implementational complexity.

G EXTRA QUALITATIVE SAMPLES

CMLM	**WMT'14 De-En**
Target	The rate of 3.1 per cent is indeed better than the previous year and is also better than in September , " however , we had hoped for more , " said Monika Felder - Bauer , acting branch manager of the Employment Agency in Sonthofen .
Teacher	Although the quota was better 3.1 better than last year and better than September , we would hoped more , " Monika - Bauer Deputy of the Labour Agency in Sonthofen .
Student	Although the quota at 3.1 % is better than last year and is also better than in September , " we would have hoped for more , " says Monika Felder - Bauer , deputy head of the Labour Agency in Sonthofen .

CMLM	**WMT'16 Ro-En**
Target	we must ask these people to learn the language , try to appropriate our values , to stop having one foot in europe and one in their home country , bringing the rest of the family including through marriages of convenience .
Teacher	let us ask these people to learn their language , try to take values , stop longer stand in europe and with one their country home origin , bringing the rest of family , including through convenience marriages .
Student	let us ask these people to learn the language , try to take over values , no longer stand in europe and with one in their home country , bringing the rest of the family , including through convenience marriages .

CMLMC	**WMT'14 De-En**
Target	Edward Snowden , the US intelligence whistleblower , has declared that he is willing to travel to Berlin to give evidence to the German parliament if the US National Security Agency and its director Keith Alexander fail to provide answers about its activities .
Teacher	Edward Snowden , the whistleblower of the US intelligence , has that he is to travel to Berlin and testify the German destag if the National Security Agency its director Keith Alexander not provide answers about their activities .
Student	Edward Snowden , the whistleblower of the US intelligence , has said that he is prepared to travel to Berlin and testify to the German destag if the American National Security Agency and its director Keith Alexander do not provide answers to their activities .

CMLMC	**WMT'16 Ro-En**
Target	during the routine control , the border policemen observed in the truck 's cab , a large travel bag , which contained personal things of many people , which is why they conducted a thorough check on the means of transport .
Teacher	at specific control , border police officers a large travel of travel the cabvan , where there were things for several people , which is why carried out thorough control over the vehicle of transport .
Student	at specific control , border police observed , in the cabin , a large travel getravel , where there were personal things for several people , which is why they carried out thorough control over the vehicle of transport .

Imputer	WMT'14 De-En
Target	German companies believe the US now poses almost as big a risk as China when it comes to industrial espionage and data theft , according to a survey published in July by EY , the consultancy .
Teacher	German companies believe that the US now poses almost as a risk as China when it to industrial espionage and data theft , according to a survey published July by consultancy firm EY in July .
Student	German companies believe that the US now poses almost as much a risk as China when it comes to industrial espionage and data theft , according to a survey published in July by the consultancy firm EY .