A. Appendix

A.1. Pytorch-Style Pseudo-Code of OAXE

Algorithm 1 Order Agnostic Cross Entropy

Input: Ground truth Y, NAT output log predictions P
bs, $length$ = Y.size()
Y = Y.repeat(1, $length$).view(bs, $length$, $length$)
$costMatrix$ = -P.gather(index=Y, dim=2)
for $i = 0$ **to** bs **do**
  $bestMatch$[i] = HungarianMatch($costMatrix$[i])
end for
Return:$costMatrix$.gather(index=$bestMatch$, dim=2)

A.2. Details for Synthetic Ordering Experiment

Distribution The categorical distributions for synthetic ordering experiment are randomly sampled from the Dirichlet Distribution, we list these distributions at following:

- 2 Modes: [0.53, 0.47]
- 3 Modes: [0.23, 0.44, 0.33]
- 4 Modes: [0.17, 0.28, 0.14, 0.41]
- 5 Modes: [0.14, 0.25, 0.13, 0.39, 0.09]

Inference Due to this experiment only cares about the multimodality of word orders, to erase the error caused by length predictor, for all CMLM models we decoding with the reference target length.

A.3. Impact of Pre-Trained XE-Based NAT Models

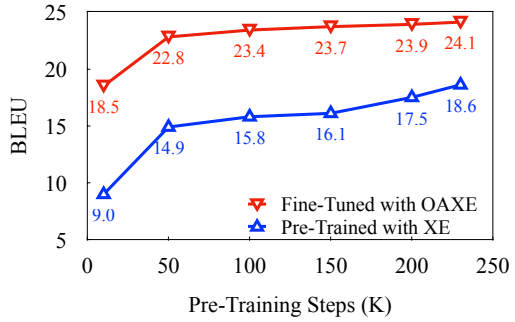

Figure 6. Impact of pre-trained NAT models with different training steps on the WMT14 En-De validation set.

Figure 6 shows the impact of different XE pre-training steps on the WMT14 En-De validation set. For each pre-trained model, we fine-tune with OAXE for 10 more epochs. OAXE consistently and significantly improves performance at all steps, demonstrating its robustness. Encouragingly, given a pre-trained NAT at early stage (e.g., 15.8 at 100K steps), fine-tuning with OAXE achieves 23.4 BLEU with 11K more training steps, which outperforms both the aligned and vanilla XEs with fewer training steps (111K vs 230K).

A.4. Different Masking Objectives

Training Objective		WMT14	
Input Tokens	**Loss Functions**	**En-De**	**De-En**
Unobserved	*All Tokens*	24.1	29.4
Partially-Observed	*All Tokens*	24.0	29.1
Partially-Observed	*Only Masks*	24.0	29.0

Table 11. Ablation study of different masking strategies for OAXE on WMT14 En⇔De validation set.

Due to the flexibility of CMLM, there are multiple different training mask strategies. However, we found our method is not sensitive to the mask strategies as shown in Table 11. The strategy *Mask All and Predict All Tokens* achieves best performance and we attribute it to the consistency of training and inference for purely non-autoregressive translation. Due to its simpleness and effectiveness, we choose it as the default setting.

A.5. Results of Sequence Lengths on Full Test Sets

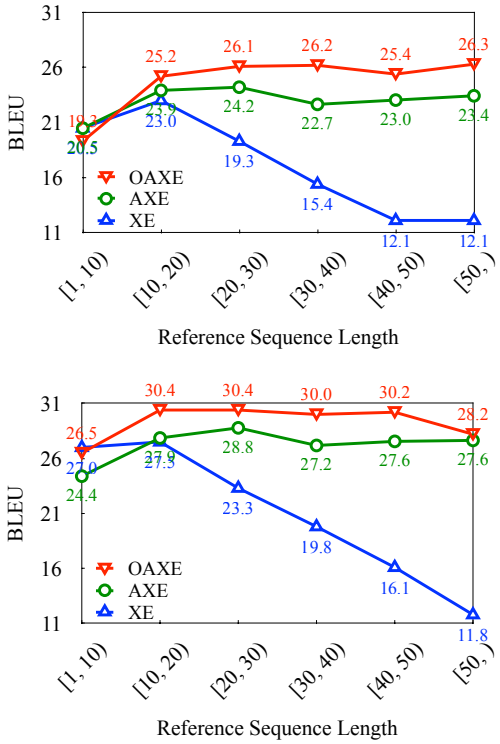

Figure 7. Performance of different sequence lengths on the distilled data for WMT14 En-De (upper panel) and De-En (bottom panel).