A. Effect of Dropout

In Section 5.1 we observed that it is crucial to turn off dropout during the computation of responsibilities (E-step) to avoid model collapse, see Table 1. In this section, we further investigate the impact of dropout on the responsibility computation, using hMup as an example.

We speculate that dropout noise weakens the dependency on the latent variable, causing the hard E-step to select among the latent values at random. This prevents different latent states from specializing and ultimately causes the model to ignore them.

To test this hypothesis, we show in Figure 4 the effect of dropout noise on the E-step at the beginning of training (i.e., with a randomly initialized model). On the y-axis we plot how often the optimal value of z changes after applying dropout with different rates. We see that as we increase the dropout probability, the optimal value of z is quickly corrupted—even with a small dropout probability of 0.1 we observe a 42% chance that the optimal assignment of z changes.

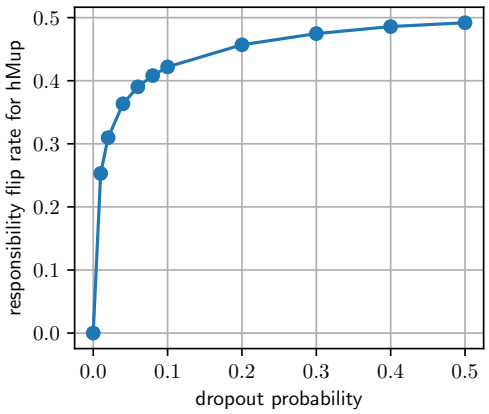

Figure 4. The effect of dropout on the E-step at the beginning of training. On the x-axis the dropout rate, and on the y-axis the fraction of times that the (hard) responsibility assignment is changed when dropout noise is turned on. Even a small amount of dropout noise causes excessive randomness of responsibility assignment, which in turn makes the model ignore the latent variable. It is therefore important to turn off dropout when estimating responsibilities. Experiments are performed on the WMT'17 En-De dataset with 2 latent categories ($K = 2$) and the "base" Transformer architecture.

B. Another Diversity Metric

In addition to Pairwise-BLEU, we also consider another metric to evaluate the diversity of a set of hypotheses, the **reference coverage**.

We pair each hypothesis to its best matching reference (breaking ties randomly), count how many distinct references are matched to at least one hypothesis, and average this count over all sentences in the test set. A low coverage number indicates that all hypotheses are close to a few references. Instead, we would like a diverse set that covers most of the references.

Compared to Pairwise-BLEU, this metric offers more intuitive numerical values, as it ranges between 1 and the total number of available references. In the next section, we will report both values for completeness.

C. Detailed Results

Table 5 compares two well performing mixture model variants, namely (online, shared) hMup and sMup, to several baselines on the three banchmark datasets we have considered in this work, expanding on the results reported in Table 2 and Figure 3 of the main paper. Once again, mixture models offer a good trade-off between translation quality and diversity.

In Table 6 we compare different approaches for generating diverse translations on the WMT'17 En-De dataset. We additionally compare each approach as we vary the number of desired translations (K) (see also Figure 3, *left*). We observe that sampling produces diverse but low quality outputs. We can improve translation quality by restricting sampling to the top-k candidates at each output position ($k = 2$ performed best), but translation quality is still worse than hMup. Beam search produces the highest quality outputs, but with low diversity. Diverse beam search provides a reasonable balance between diversity and translation quality, but hMup produces even more diverse and better quality translations. Finally, except for unrestricted sampling, hMup covers the largest number of references among all the approaches evaluated.

We conclude with Table 7 which shows the values used to generate Figure 2, together with other metrics, such as corpus level BLEU with hypotheses generated by a fixed latent variable state throughout the whole test set. This metric is useful to detect models affected by degeneracy D1, as there will be states that yield very low corpus level BLEU because they rarely generate good hypotheses.

	Pairwise-BLEU			BLEU			#refs covered		
	en-de	en-fr	zh-en	en-de	en-fr	zh-en	en-de	en-fr	zh-en
Sampling	24.1	32.0	48.2	37.8	46.5	19.5	4.6	4.3	1.5
Beam	73.0	77.1	83.4	69.9	79.8	33.9	3.1	3.2	1.3
Diverse beam	53.7	64.9	66.5	60.0	72.5	31.6	3.7	3.5	1.4
sMup (He et al., 2018)	68.9	80.4	60.9	68.1	79.6	32.8	2.9	2.7	1.4
hMup	50.2	64.0	51.6	63.8	74.6	31.9	4.0	3.7	1.6
Human	35.5	46.5	25.3	69.6	76.9	38.0	-	-	-

Table 5. Results on three WMT datasets. Extended version of Table 2, including the results of another mixture model sMup (He et al., 2018) and the reference coverage metric. hMup and sMup provide different trade-offs between quality and diversity: the former is more diverse (lower Pairwise-BLEU), while the latter gives higher translation quality (BLEU).

	Pairwise-BLEU			BLEU			#refs covered		
	$K=5$	10	20	5	10	20	5	10	20
Sampling	31.6	24.1	21.2	37.8	37.8	37.9	3.1	4.6	6.2
Biased sampling (top-2)	49.3	47.8	46.7	59.7	60.0	60.4	2.7	3.7	4.7
Beam	77.1	73.0	69.1	70.7	69.9	68.7	2.3	3.1	4.0
Diverse beam	59.8	53.7	49.7	63.4	60.0	57.7	2.5	3.7	4.8
hMup	54.2	50.2	47.1	64.9	63.8	62.3	2.8	4.0	5.3

Table 6. Results on the WMT'17 En-De dataset with various numbers of generations (K). We compare: multinomial sampling (`Sampling`); sampling restricted to the top-k candidates at each step (`Biased sampling (top-2)`; $k=2$ performed best); beam search with varying beam widths (`Beam`); diverse beam search (Vijayakumar et al., 2018) with varying number of outputs (`Diverse beam`; note that the number of groups G and diversity strength are tuned separately for each value of K); and the hMup mixture model with K components (`hMup`).

schedule	parameterization	loss	BLEU per latent $z=1$	2	3	Pairwise-BLEU	BLEU	#refs covered
online	shared	sMlp	25.9	24.9	22.6	57.8 (4.0)	64.4 (1.0)	1.9
online	shared	sMup	25.8	25.2	22.8	61.6 (1.3)	64.2 (0.4)	1.9
online	shared	hMlp	25.5	22.6	21.5	47.8 (0.6)	60.3 (0.3)	2.1
online	shared	hMup	25.6	24.4	21.3	53.1 (1.2)	62.6 (0.6)	2.1
online	independent	sMlp	25.5	0.0	0.0	23.8 (23.8)	13.4 (9.6)	2.6
online	independent	sMup	25.8	0.6	0.0	3.5 (2.9)	17.8 (8.1)	2.6
online	independent	hMlp	26.1	0.0	0.0	0.0 (0.0)	3.7 (0.0)	2.7
online	independent	hMup	25.7	0.2	0.0	0.1 (0.1)	18.7 (6.0)	2.4
offline	shared	sMlp	25.8	25.7	25.6	91.7 (6.8)	66.8 (0.7)	1.4
offline	shared	sMup	25.7	25.5	25.3	91.7 (9.6)	66.8 (0.9)	1.4
offline	shared	hMlp	25.6	21.2	19.4	49.2 (7.3)	57.7 (5.0)	2.1
offline	shared	hMup	25.3	24.4	23.6	58.2 (1.5)	63.7 (0.6)	2.0
offline	independent	sMlp	25.7	19.4	16.4	34.4 (7.8)	53.6 (5.7)	2.2
offline	independent	sMup	25.3	23.5	22.9	48.0 (1.1)	62.3 (0.6)	2.1
offline	independent	hMlp	25.7	15.8	13.0	26.4 (9.1)	47.6 (7.0)	2.4
offline	independent	hMup	25.5	22.5	19.8	42.5 (2.6)	59.1 (1.6)	2.2

Table 7. Comparison of mixture models with different design choices on the WMT'17 En-De dataset with $K=3$ mixture components. See §3.1, §3.2, §3.3 for a detailed discussion about these model configurations. Pairwise-BLEU versus BLEU are plotted in Figure 2. Each configuration was run five times with different random seeds. We report the mean value of each metric (and for Pairwise-BLEU and BLEU also the standard deviation in parentheses). We also report corpus level BLEU w.r.t. one reference when greedily decoding the test set using a fixed latent variable state (columns labeled with $z=1,2,3$). Configurations that have values colored in red exhibit degeneracy of type D1, while configurations that have values colored in blue exhibit degeneracy of type D2.