6. Supplementary

6.1. Adding Cross Distillation Loss to Baseline

Similar to [25, 24], Yang *et al*. [32] learns to predict the next step in the recipe from a cooking video by transferring knowledge from the textual domain to visual domain. Motivated by knowledge distillation [13], they use the textual future prediction model as teacher, while training a student model on videos. They call their method cross-modal contrastive distillation (CCD). They do this process after finetuning the textual model on the text of the video dataset (YouCookII in our case). For additional comparison, we extend our baseline from the main paper by training it using CCD on YouCookII. We report results in table 5. While the baseline method benefits from CCD, it still fails to match the performance of our GEPSAN. The results indicate superiority of our generative approach over cross-modal distillation.

		ING	VERB	B1	B4	MET
Single *(S)* Prediction	Baseline	19.6	27.5	25.8	4.0	9.8
	CCD	20.8	27.0	26.4	4.2	10.0
	GEPSAN	25.6	30.8	28.9	5.8	11.8
Multiple *(M)* Predictions	Baseline$^\diamond$	32.2	34.2	35.0	5.9	13.7
	CCD$^\diamond$	33.5	34.3	36.2	6.8	14.1
	GEPSAN	**36.7**	**38.4**	**37.1**	**9.3**	**15.7**

Table 5. Future anticipation results on YouCookII Video after finetuning. CCD is our adaptation of [32] to the Baseline. $^\diamond$ We use *Nucleus sampling* [14] to achieve multiple predictions from the deterministic baseline.

6.2. Impact of k

In Fig. 4, we show BLEU4 and METEOR scores of our GEPSAN and the baseline with increasing value of k. Observe that the baseline fails to match the performance of GEPSAN, suggesting that simply increasing k is not sufficient to improve performance if diversity is not meaningful. Higher performance of our model across k shows that it indeed produces diverse yet meaningful predictions. Also the improvements with increasing k plateaus early, indicating that GEPSAN predicts steps highly relevant to the GT.

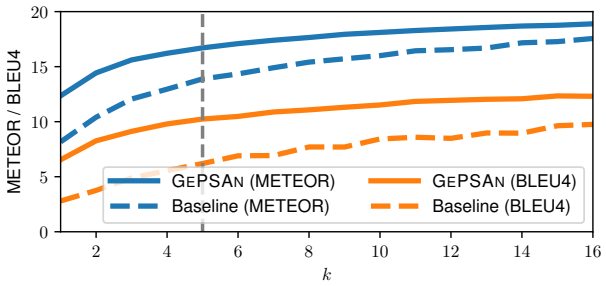

Figure 4. METEOR and BLEU4 with increasing number of sampled predictions (k) on YouCookII videos after finetuning.

6.3. Additional results with different evaluation methodologies

In Tables 1 and 4 (main paper), we report our and the baseline results for BLEU1 (B1), BLEU4 (B4) and METEOR (MET) metrics computed using the standard corpus-level formula which uses the micro-averaged statistics before computing the corpus-level BLEU and METEOR scores [21]. However, the baseline results were originally reported by macro-averaging the sentence-level metrics [24]. Here, in Table 6, we report our results, as well as the reproduced baseline results, using both micro- and macro-averaging. We also present the original macro-averaged baseline results as reported in [24]. We observe that the baseline macro-averaging results reproduced by us are similar to the ones reported in the original paper. Also, the micro-averaged results follow the same trend as the macro-averaged ones.

6.4. Results on the YouCookII standard splits

In Tables 1 and 2 (main paper), we report results for the YouCookII splits proposed by [24], where each split represents a different set of dishes out of the 89 dishes (no overlapping dishes between the different splits). This allowed for comparing

Model	ING		VERB		B1			B4			MET		
	Micro	*Micro* [24]	*Micro*	*Micro* [24]	*Micro*	*Macro*	*Macro* [24]	*Micro*	*Macro*	*Macro* [24]	*Micro*	*Macro*	*Macro* [24]
YouCookII Video (Unseen Split)													
BASELINE *(S)*	16.8	17.8	26.9	23.1	25.1	22.4	20.6	3.1	0.6	0.84	9.2	9.9	9.5
GEPSAN *(S)*	21.5	-	29.9	-	27.6	25.6	-	4.8	1.4	-	10.8	12.0	-
BASELINE *(M)*	27.8	-	31.6	-	33.1	28.8	-	4.4	0.8	-	12.2	13.2	-
GEPSAN *(M)*	**31.6**		**37.8**	-	**35.6**	**33.0**	-	**7.9**	**2.6**	-	**14.5**	**16.0**	-
YouCookII Video (Seen Split)													
BASELINE *(S)*	19.6	20.9	27.5	24.8	25.8	22.9	22.1	4.0	1.0	1.2	9.8	10.6	10.7
GEPSAN *(S)*	25.6	-	30.8	-	28.9	26.8	-	5.8	2.2	-	11.8	13.4	-
BASELINE *(M)*	32.2	-	34.2	-	35.0	30.9	-	5.9	1.5	-	13.7	14.8	-
GEPSAN *(M)*	**36.7**	-	**38.4**	-	**37.1**	**35.0**	-	**9.3**	**3.9**	-	**15.7**	**17.7**	-
Recipe1M+													
BASELINE *(S)*	27.0	33.5	29.4	26.7	24.1	22.1	22.8	7.8	4.1	4.4	11.3	13.4	13.7
GEPSAN *(S)*	27.2	-	28.5	-	25.9	21.1	-	7.5	3.4	-	11.2	12.3	-
BASELINE *(M)*	34.7	-	34.6	-	31.7	28.5	-	9.4	4.9	-	14.2	16.7	-
GEPSAN *(M)*	**37.2**	-	**36.2**	-	**32.2**	**29.0**	-	**10.7**	**5.6**	-	**14.6**	**16.9**	-

Table 6. We reproduce and compare the various results corresponding to Tables 1 and 4 computed using micro-averaging (*Micro*) [21] vs macro-averaging (*Macro*) [24] of the metrics. When available, we also present the exact numbers reported in the original paper (*Macro* [24]). Note, Sener *et al*. [24] report results computed using macro-averaging (*Macro*) only.

the setups where the model has never seen a specific dish before (Unseen Split) vs the setup where the model has seen the dish prepared using other (different) videos (Seen Split). The results were obtained by applying cross-validation on each of the four splits. Here, in Table 7, we report results on the original training/validation splits of YouCookII, where the videos are randomly chosen without taking into account which dish they belong to, and hence we can see that it mostly resembles the (Seen Split) setup.

Model	ING		VERB		B1			B4			MET		
	Micro	*Micro* [24]	*Micro*	*Micro* [24]	*Micro*	*Macro*	*Macro* [24]	*Micro*	*Macro*	*Macro* [24]	*Micro*	*Macro*	*Macro* [24]
BASELINE *(S)*	19.7	21.4	27.3	27.6	26.2	23.0	23.7	3.9	1.1	1.7	9.9	10.9	11.5
GEPSAN *(S)*	25.7	-	32.2	-	30.0	27.3	-	6.4	2.3	-	12.2	13.8	-
BASELINE *(M)*	33.7	-	34.1	-	36.2	31.5	-	6.4	1.7	-	14.1	15.2	-
GEPSAN *(M)*	37.2	-	40.5	-	38.4	35.8	-	9.8	4.0	-	16.3	18.3	-

Table 7. Future anticipation results on YouCookII Video using the original train/val splits [40]. These are the validation results obtained by finetuning different models on the original training split of the YouCookII dataset.