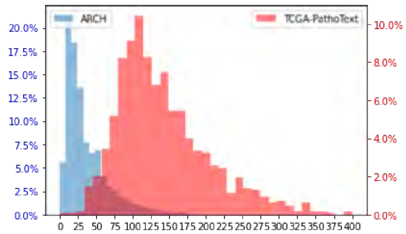
Supplementary

No Author Given

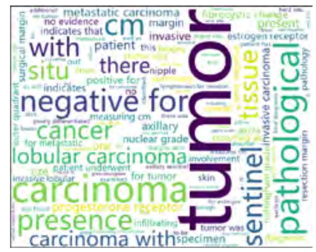
No Institute Given

Subset	Number of pairs
BRCA	1061
KIRC	513
THCA	506
UCEC	504
LUSC	478
LUAD	476
LGG	466
HNSC	450
COAD	450
SKCM	431
STAD	416
PRAD	403
BLCA	386
LIHC	365
CESC	269
SARC	247
PCPG	175
ESCA	156
TGCT	144
THYM	121
OV	106
KICH	94
UVM	80
MESO	75
UCS	57
ACC	56
DLBC	44
CHOL	39

Table 1. Composition of TCGA-PathoText



(a)



(b)

Fig. 1. (a) Histogram of text lengths. It shows that TCGA-PathoText includes longer pathology reports compared to ARCH which only describes small patches. (b) Word cloud showing 100 most frequent tokens.

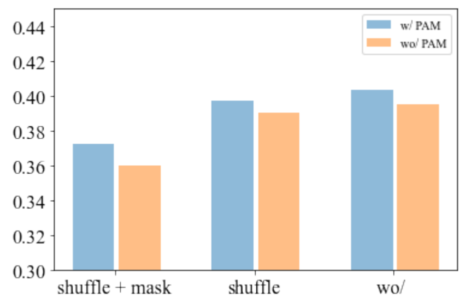


Fig. 2. Effect of patch shuffling. We add noise into the spatial context by shuffling the patches in the training stage. And the test data is not shuffled or masked. We still use BLEU-1 to measure the generation performance. "shuffle + mask" represents that we shuffle and mask the tokens at the same time.

	BLEU-4	BLEU-1
Single Layer	0.092	0.377
2d sin-cos	0.093	0.380
3×3+5×5+7×7	0.096	0.385
3×3	0.095	0.381
7×7	0.090	0.383
w/o	0.089	0.395
Ours	**0.117**	**0.403**

Table 2. Effect of Position-aware Module with different structures. 'Single Layer' represents we only insert the PAM after the final encoder block. Compared with the model without any position-aware modules, we can see that PAM can consistently improve the performance whatever the structure is. And applying convolutional layers with only one type of kernel improves the diagnosis not as well as the combination of different kernels.