

Supplementary Material for "Masked Vision-Language Transformers for Scene Text Recognition"

			
MVLT: brown MVLT*: brown GT: brown	MVLT: celebrating MVLT*: celebrating GT: celebrating	MVLT: currency MVLT*: currency GT: currency	MVLT: beaut MVLT*: beaute GT: beaute
			
MVLT: buffet MVLT*: buffet GT: buffet	MVLT: louisiana MVLT*: louisiana GT: louisiana	MVLT: araldi9930 MVLT*: arald11930 GT: arald11930	MVLT: gruha MVLT*: gruha GT: gruha
			
MVLT: designstockz MVLT*: designstockz GT: designstockz	MVLT: dreamer MVLT*: dreamer GT: dreamer	MVLT: diviniion MVLT*: division GT: division	MVLT: kingfisher MVLT*: kingfisher GT: kingfisher
			
MVLT: meant MVLT*: meant GT: meant	MVLT: ka MVLT*: kia GT: kia	MVLT: mateer MVLT*: manchester GT: manchester	MVLT: tournesol MVLT*: tournesol GT: tournesol
			
MVLT: garage MVLT*: garage GT: garage	MVLT: watchmen MVLT*: watchmen GT: watchmen	MVLT: mint MVLT*: mint GT: mint	MVLT: shining MVLT*: shining GT: shining
			
MVLT: gola MVLT*: cola GT: cola	MVLT: space MVLT*: space GT: space	MVLT: anaheim MVLT*: anaheim GT: anaheim	MVLT: venus MVLT*: venus GT: venus
			
MVLT: wwwnnntoppiggnccm MVLT*: wwwnonstopsignscom GT: wwwnonstopsignscom	MVLT: wwwshutterstockcom MVLT*: wwwshutterstockcom GT: wwwshutterstockcom	MVLT: wwwposterpluscouk MVLT*: wwwposterpluscouk GT: wwwposterpluscouk	MVLT: wwwloudbillboardsscm MVLT*: wwwloudbillboardscom GT: wwwloudbillboardscom

Figure 1: Visualization of fine-tuned MVLT and MVLT*. For each example, we show the STR results of MVLT, MVLT*, and the ground-truth. The wrong predictions are shown in red. The images are from test datasets.



Figure 2: Visualization of pretrained MVLT and MVLT*. For each example, we show the masked image (left), the image reconstruction result of *decoder*₂ of the pretrained MVLT (mid-left), the image reconstruction result of *decoder*₂ of the pretrained MVLT* (mid-right), and the ground truth (right). The images are from test datasets.