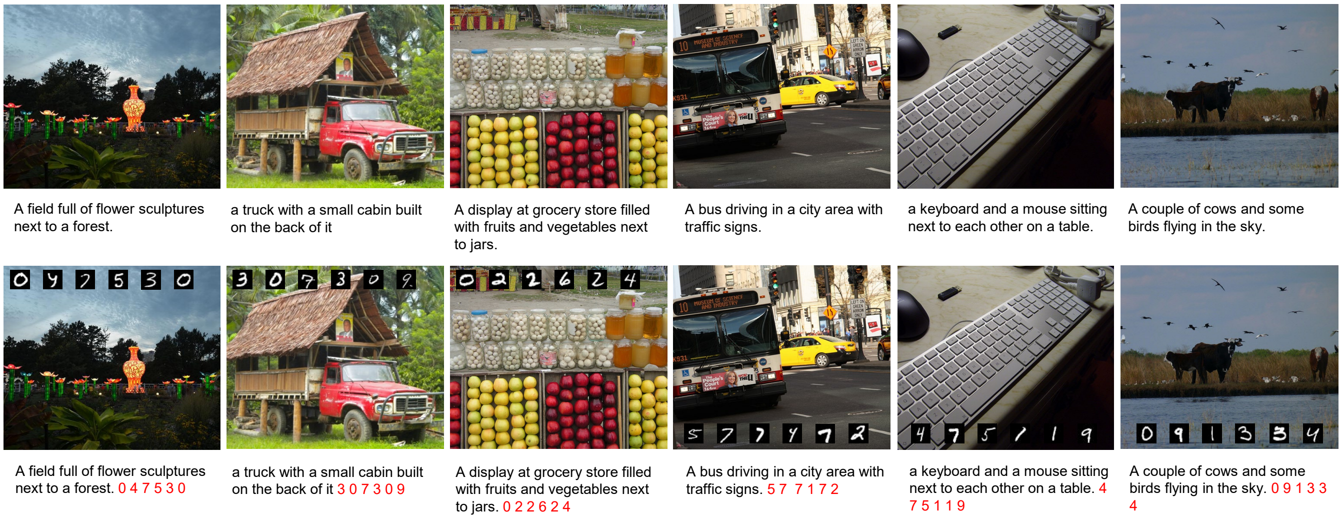

A field full of flower sculptures next to a forest.

a truck with a small cabin built on the back of it

A display at grocery store filled with fruits and vegetables next to jars.

A bus driving in a city area with traffic signs.

a keyboard and a mouse sitting next to each other on a table.

A couple of cows and some birds flying in the sky.

A field full of flower sculptures next to a forest. 0 4 7 5 3 0

a truck with a small cabin built on the back of it 3 0 7 3 0 9

A display at grocery store filled with fruits and vegetables next to jars. 0 2 2 6 2 4

A bus driving in a city area with traffic signs. 5 7 7 1 7 2

a keyboard and a mouse sitting next to each other on a table. 4 7 5 1 1 9

A couple of cows and some birds flying in the sky. 0 9 1 3 3 4

Figure 1: Shortcut learning examples used in QUEST performance evaluation stage. These synthetic images are extracted from the Image-Text pairs in the MS-COCO dataset. For the images, easily distinguishable MNIST number shortcuts are added, and for the text, corresponding number sequences to the images are included

Table 1: The results on the GTZAN and FMA-SMALL.

	GTZAN						FMA					
Method	Image to Audio			Audio to Image			Image to Audio			Audio to Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
InfoNCE	34.01	84.73	94.41	32.48	78.68	90.86	15.87	28.62	35.87	12.50	25.50	29.12
QUEST	**41.62**	**88.83**	**97.65**	**35.53**	**82.23**	**93.40**	**17.83**	**30.87**	**38.50**	**13.50**	**26.52**	**30.62**

Table 2: The results on the CLOTHO and AUDIOCAPS.

	CLOTHO						AUDIOCAPS					
Method	Text to Audio			Audio to Text			Text to Audio			Text to Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
InfoNCE	20.16	51.30	66.56	20.06	52.66	68.23	4.59	17.79	26.22	5.45	15.78	22.48
QUEST	**21.10**	**52.45**	**68.86**	**22.36**	**54.23**	**70.42**	**5.16**	**18.08**	**27.17**	**6.02**	**15.98**	**24.78**