Figure A: Examples of **unconditional image generation** on CelebA-HQ based on VQ-GAN+LG.

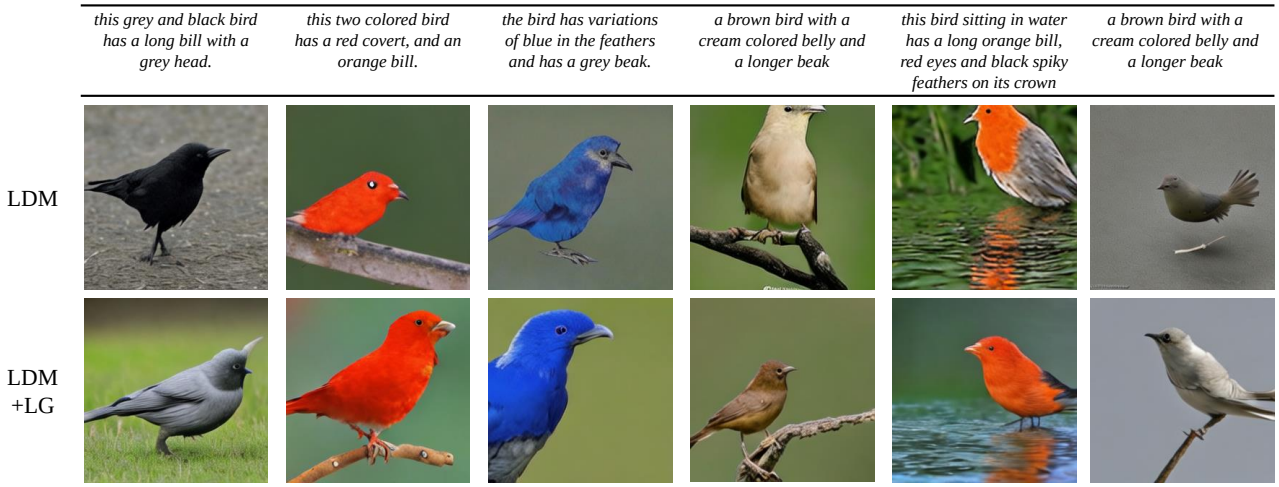

Figure B: Examples of **text-to-image generation** on CUB-200.

Model	Image Reconstruction on MS-COCO FID↓	VQA on MS-COCO Accuracy↑
VQCT	9.82	40.42
VQCT+LG	**9.57**	**40.64**

Table 1: Comparison of **reconstruction and VQA** on VQCT and VQCT+LG on the MS-COCO dataset

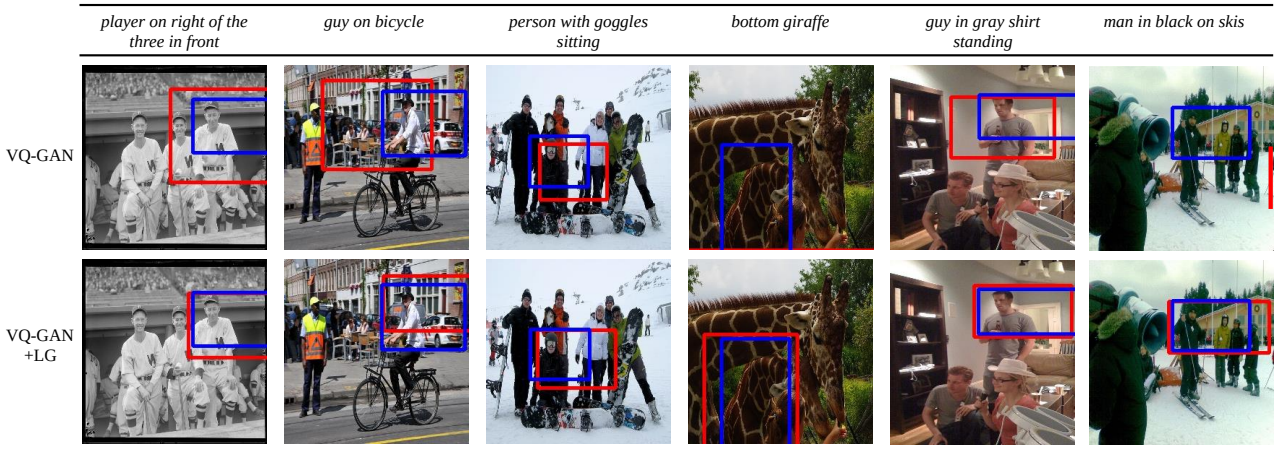

Figure C: Examples of **visual grounding** on refcoco. Blue boxes are the ground-truth, red boxes are the predictions of the model.