



Figure 4. Selected examples of **instructed zero-shot image-to-text generation** using a BLIP-2 model w/ ViT-g and FlanT5<sub>XXL</sub>, where it