Though image-to-sequence generation models have become overwhelmingly popular in human-computer communications, they suffer from strongly favoring sa