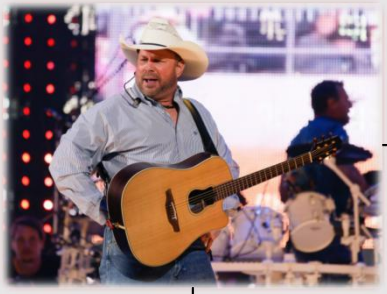
Training Data Generation

(a) Generate caption and tags:



BLIP2

"a man in a cowboy hat and jeans is holding an acoustic guitar"

Spacy

(b) Generate boxs and masks:

Grounding DINO + SAM

man

hat

jeans

guitar

(c) Compose structured data:

Caption:
a man in a cowboy hat and jeans is holding an acoustic guitar