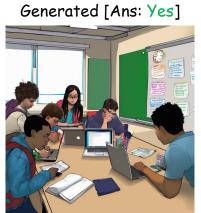
Text Guidance: Are there six people? Reference [Ans: Yes]



I2T2I SDXL

✓ Diversity X Fidelity

(a) Numerical Attribute





(d) Spatial Relationship

Reference Image

Reference [Ans: No]

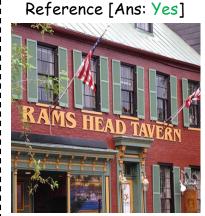


Generated [Ans: No]

(b) Physical Attribute (Color)

Is there a red hat?

Is this a general store outdoor?





Generated [Ans: Yes]







X Diversity ✓ Fidelity

Is there a bottle?



Generated [Ans: Yes]

(c) Object Existence

Is person swinging or holding a tennis racket?



Generated [Ans: swing]

(f) Human-Object Interaction Reasoning

(g) Diversity/Fidelity Comparison

Textual Inversion



✓ Diversity X Fidelity

Hummingbird

 $\checkmark$  Diversity  $\checkmark$  Fidelity



Context Description from MLLM: The image shows a city street scene with a clock mounted on a building. The clock is positioned above the street level, where pedestrians are walking. The street is lined with shops and buildings, and there are people walking on the sidewalk. The clock is clearly visible and serves as a prominent feature in the image.