Q:Is the word "DRINKS"placed below "COLD" or above it?

(a) Target-Driven(ours)

(b) Text to Image Attention

(c) Diversity

(d) Cls-token Attention

Input Image

A:"Below" ✓

A:"Below" ✓

A:"Above" ✗

A:"Above" ✗

A:"Above" ✗

Q:Output a single digit only: how many animals are on the bed?

(a) Target-Driven(ours)

(b) Text to Image Attention

(c) Diversity

(d) Cls-token Attention

Input Image

A:"3" ✓

A:"3" ✓

A:"2" ✗

A:"2" ✗

A:"2" ✗

Q:Output a single digit only: how many small condiment cups contain dark red jam?

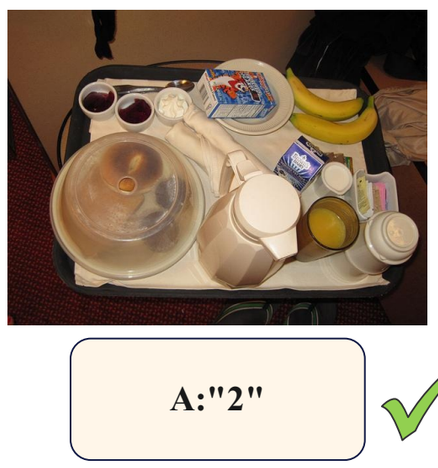
(a) Target-Driven(ours)

(b) Text to Image Attention

(c) Diversity

(d) Cls-token Attention

Input Image



A:"2" ✓

A:"2" ✓

A:"1" ✗

A:"1" ✗

A:"1" ✗

Q:Answer with exactly one token (YES/NO): is there a person standing on the platform?

(a) Target-Driven(ours)

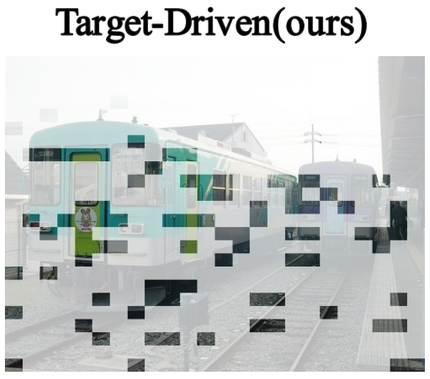
(b) Text to Image Attention

(c) Diversity

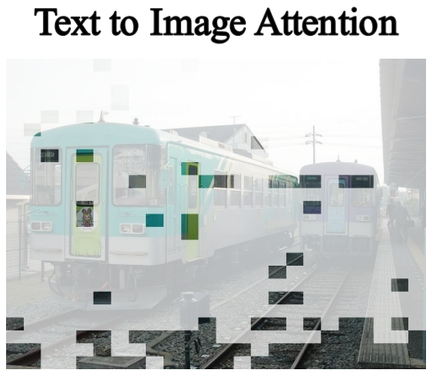
(d) Cls-token Attention

Input Image

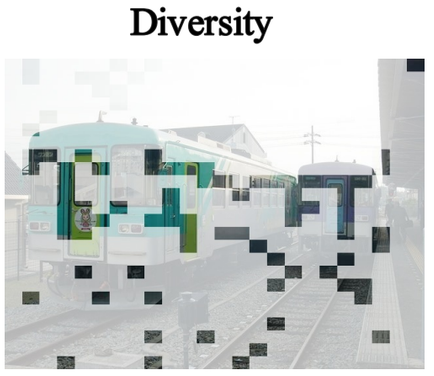
A:"Yes" ✓



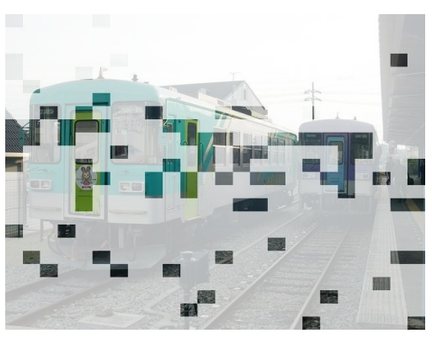
A:"Yes" ✓



A:"No" ✗



A:"No" ✗



A:"No" ✗

Q:Output a single digit only: how many balloons are tied to the stand near the back desk?

(a) Target-Driven(ours)

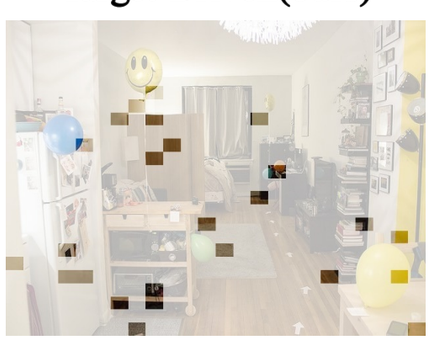
(b) Text to Image Attention

(c) Diversity

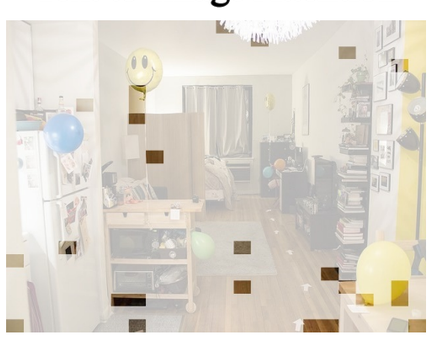
(d) Cls-token Attention

Input Image

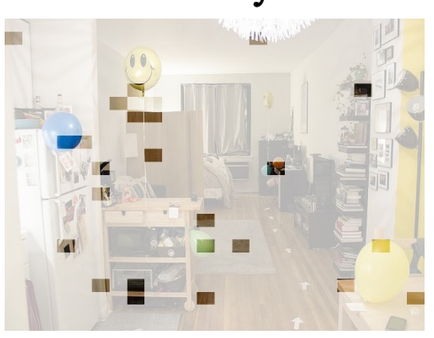
A:"3" ✓



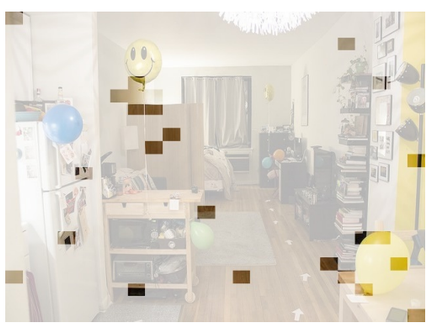
A:"3" ✓



A:"0" ✗



A:"2" ✗



A:"2" ✗