A. Examples of SpatialBench

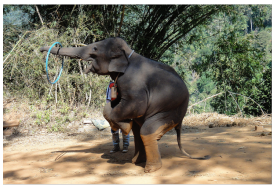
In Fig. 1 and 2, we show outdoor and indoor examples from the SpatialBench, respectively. Our benchmark constructs misleading hard negative captions to evaluate the model's ability to understand various spatial concept, including depth, orientation, spatial state, object relationships and object size, etc.

B. Template for Caption Perturbation

We use templates to perturb the spatial description in captions, and the detailed replacing templates are presented in Tab. 1. For spatial phrases, we replace them with antonyms randomly selected from their corresponding options. For those captions where no spatial description is found, we randomly select a caption from another image as its negative example to maintain the consistency of the method.

While we rely on the simple template to substitute spatial descriptions in captions, this approach demonstrated empirical effectiveness in experiments and greatly enhanced the CLIP model's understanding of spatial concepts.

A. Trains running on this track will go straight.
B. Trains running on this track will turn to the right.
C. Trains running on this track will turn to the left.
D. Trains running on this track will have a sharp U-turn.



A. The elephant is lying down.
B. The elephant is standing on its all four legs.
C. The elephant is standing on its front legs.
D. The elephant is standing on its hind legs.

A. All three people are standing on the snow.
B. One person is standing and two others are jumping in the air.
C. Two people are standing and one is jumping in the air.
D. All three people are jumping in the air."

A. The car closest to the camera is facing the same direction as the truck next to it.
B. The car closest to the camera is facing the opposite direction as the truck next to it.

A. The two motorbikes are facing the same direction.
B. The two motorbikes are facing opposite directions.

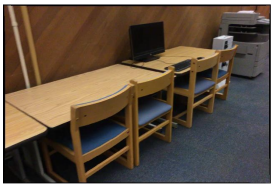
A. The four zebras in the picture look to the left.
B. The four zebras in the picture look to the right.
C. The four zebras look towards the camera.
D. The four zebras in the picture face away from the camera.

A. The bench on the left is larger than bench on the right.
B. The bench on the left is the same size as bench on the right.
C. The bench on the left is smaller than bench on the right.

A. The bird on the left is furthest away from the camera.
B. The bird on the right is farthest from the camera.
C. The bird in the middle is furthest away from the camera.
D. The three birds are the same distance away from the camera.

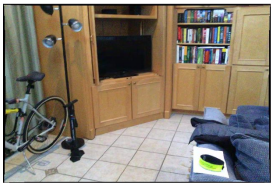
A. There are no cars behind the fire truck.
B. A car is directly behind the fire truck.
C. Two cars are directly behind the fire truck.
D. Three cars are directly behind the fire truck.

Figure 1. Examples from the outdoor subset of SpatialBench.

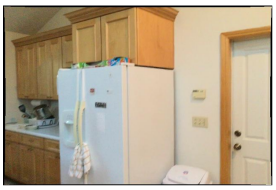


A. The computer is in front of the first chair from right to left.
B. The computer is in front of the second chair from right to left.
C. The computer is in front of the third chair from right to left.
D. The computer is in front of the fourth chair from right to left.

A. The color of the chair closest to the camera is brown.
B. The color of the chair closest to the camera is dark blue.
C. The color of the chair closest to the camera is gray.

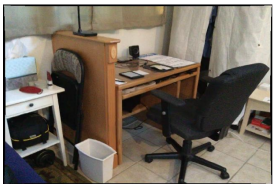


A. The television is hanging on the wall.
B. The television is hanging outside the cabinet.
C. The television is put inside the cabinet.



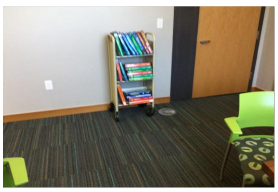
A. The refrigerator is taller than the door on the right.
B. The refrigerator is shorter than the door on the right.
C. The refrigerator is the same height than the door on the right.

A. The two desks in the middle are of same size.
B. The desk on the left is larger than the desk on the right.
C. The desk on the left is smaller than the desk on the right.

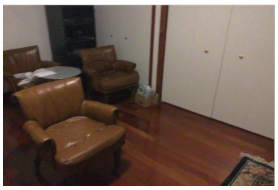


A. The white garbage can is under the wooden table.
B. The white garbage can is next to the wooden table.
C. The white garbage can is behind the wooden table.
D. The white garbage can is between the wooden table and the black chair.

A. The green plant is in the middle of the sofas.
B. The green plant is partially blocked by the sofas.
C. The green plant is in front of the sofas.
D. The green plant is behind the sofas at a distance.



A. The book shelf is in front of the door.
B. The book shelf is in the center of the room.
C. The book shelf is leaning against the door.
D. The book shelf is leaning against the wall.



A. All three chairs are facing in the same direction.
B. One chair is facing in different direction than the other two.
C. All three chairs are facing in different direction.

Figure 2. Examples from the indoor subset of SpatialBench.

Phrases	Replaced by	Antonym Options
“above”	$\Longrightarrow$	[“below”,“under”,“beneath”,“down”,“lower”]
“below”	$\Longrightarrow$	[“above”, “over”,“higher”,“up”,“top”]
“under”	$\Longrightarrow$	[“above”,“over”, “higher”,“up”, “top”]
“in front of”	$\Longrightarrow$	[“behind”,“after”,“back”,“rear”,“following”]
“behind”	$\Longrightarrow$	[“in front of”,“ahead”,“before”,“leading”,“fore”]
“to the left of”	$\Longrightarrow$	[“to the right of”, “right”, “east”, “on the right side of”, “opposite”]
“to the right of”	$\Longrightarrow$	[“to the left of”, “left”, “west”, “on the left side of”, “opposite”]
“on the left”	$\Longrightarrow$	[“on the right”,“on the middle”]
“on the right”	$\Longrightarrow$	[“on the left”,“on the middle”]
“near”	$\Longrightarrow$	[“far from”,“distant”,“remote”,“away”,“separate”]
“far from”	$\Longrightarrow$	[“near”,“close”,“adjacent”,“next to”,“together”]
“next to”	$\Longrightarrow$	[“far from”, “away”,“distant”,“separate”,“apart”]
“between”	$\Longrightarrow$	[“outside”, “beyond”“around”,“among”,“surrounding”]
“inside”	$\Longrightarrow$	[“outside”, “outdoors”, “exterior”,“beyond”, “away from”]
“outside”	$\Longrightarrow$	[“inside”,“within”,“interior”,“enclosed”,“contained”]
“on top of”	$\Longrightarrow$	[“underneath”,“below”, “beneath”, “down”,“lower”]
“underneath”	$\Longrightarrow$	[“on top of”,“above”, “over”, “higher”, “up”]
“at the center of”	$\Longrightarrow$	[“on the edge of”, “periphery”, “border”, “outside”, “fringe”]
“on the edge of”	$\Longrightarrow$	[“at the center of”,“middle”,“core”,“inside”, “interior”]
“across from”	$\Longrightarrow$	[“adjacent to”,“next to”,“beside”,“alongside”,“near”]
“alongside”	$\Longrightarrow$	[“across from”,“opposite”,“far from”,“away from”,“distant”]
“surrounding”	$\Longrightarrow$	[“enclosed by”,“inside”,“within”,“contained”,“centered”]
“enclosed by”	$\Longrightarrow$	[“surrounding”,“outside”,“beyond”,“exterior”,“outdoors”]
“adjacent to”	$\Longrightarrow$	[“far from”,“distant”,“separate”,“away from”,“remote”]
“in the vicinity of”	$\Longrightarrow$	[“far from”,“distant”,“remote”,“away”, “separate”]
“on the surface of”	$\Longrightarrow$	[“beneath”,“under”,“below”,“down”, “lower”]
“beneath”	$\Longrightarrow$	[“on top of”,“above”,“over”,“higher”,“up”]
“in the background of”	$\Longrightarrow$	[“in the foreground of”, “front of”, “before”, “leading”]
“in the foreground of”	$\Longrightarrow$	[“in the background of”, “behind”, “after”, “rear”]
“surrounded by”	$\Longrightarrow$	[“isolated from”, “away from”,“distant from”,“separate from”]
“in the middle of”	$\Longrightarrow$	[“on the edge of”,“outside”,“beyond”,“periphery”]
“next to”	$\Longrightarrow$	[“far from”,“away from”,“distant from”,“separate from”]
“over”	$\Longrightarrow$	[“under”,“beneath”, “below”,“down”, “lower”]
“big”	$\Longrightarrow$	[“small”,“tiny”,“the same size”]
“large”	$\Longrightarrow$	[“small”,“tiny”,“the same size”]
“small”	$\Longrightarrow$	[“large”,“big”,“the same size”]
“short”	$\Longrightarrow$	[“tall”,“the same height”]
“low”	$\Longrightarrow$	[“tall”,“the same height”]
“tall”	$\Longrightarrow$	[“short”,“low”,“the same height”]
“bigger”	$\Longrightarrow$	[“smaller”,“the same size”]
“larger”	$\Longrightarrow$	[“smaller”,“the same size”]
“smaller”	$\Longrightarrow$	[“larger”,“bigger”,“the same size”]
“shorter”	$\Longrightarrow$	[“taller”,“the same height”]
“lower”	$\Longrightarrow$	[“taller”,“the same height”]
“taller”	$\Longrightarrow$	[“shorter”,“lower”,“the same height”]
“the same height”	$\Longrightarrow$	[“shorter”,“lower”,“taller”]
“the same size”	$\Longrightarrow$	[“larger”,“smaller”,“bigger”,“the bigest”,“the largest”,“the smallest”]
“the biggest”	$\Longrightarrow$	[“the smallest”,“the same size”]
“the largest”	$\Longrightarrow$	[“the smallest”,“the same size”]
“the smallest”	$\Longrightarrow$	[“the biggest”,“the laggest”,“the same size”]
“parallel”	$\Longrightarrow$	[“perpendicular”,“intersecting”,“angle with”]
“perpendicular”	$\Longrightarrow$	[“parallel”,“intersecting”,“angle with”]
“straight line”	$\Longrightarrow$	[“circle”,“curve”,“triangle”,“rectangle”]
“curve”	$\Longrightarrow$	[“circle”,“straight line”,“triangle”,“rectangle”]
“on the edge of”	$\Longrightarrow$	[“in”,“on”,“above”]

Table 1. Template for Caption Perturbation. For spatial phrases, we replace them with antonyms randomly selected from their corresponding options.