

## Text Prompts ['What are the key elements in

this picture?',
'What does the bus screen show?

'Is there any plants in the picture?',

'Is the bus full of people?',
'Is the bus in a rural place or in

the city?',
'Is the line of sight blocked?',

'What is the color of the building behind the bus?', 'Could you please tell me about

the main components of the picture?']



Vision Tokens