Reference – Shown-as-Correct Answers

Q1

Answer presented as correct: A) Mapping textual phrases to image regions■(

Reason: Involves the process of establishing a visual connection to textual elements within an image.

Q2

Answer presented as correct: D) Reinforcement learning

Reason: Utilizes a trial-and-error approach to learn optimal actions through the environment's feedback.

Q3

Answer presented as correct: A) Concatenating visual patches and word tokens into a single Transformer stream■(

Reason: Facilitates seamless integration by combining visual data representation with textual context.

Q4

Answer presented as correct: B) A CNN classification head■(

Reason: A CNN classification head interprets visual content for generating descriptive captions.

Q5

Answer presented as correct: D) High memory consumption

Reason: Critical for real-time processing, thus leading to enhanced performance despite potential drawbacks.

Q6

Answer presented as correct: B) Using retrieval-based pipelining■(

Reason: Utilizing a sequential processing approach to handle different types of information efficiently.

Q7

Answer presented as correct: B) FID■(

Reason: Measures the fidelity and quality of generated images through a perceptual distance calculation.