Reference – Shown-as-Correct Answers

Q1

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

Reason: Involves determining the spatial relationships between textual descriptions and visual elements within an image.

Q2

Answer presented as correct: D) Reinforcement learning

Reason: Utilizes a trial-and-error approach to maximize expected cumulative rewards for choosing optimal actions.

Q3

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

Reason: Efficiently integrates different modalities with improved model interpretability.

Q4

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

Reason: Responsible for identifying features in images that aid in caption creation.

Q5

Answer presented as correct: B) Table normalization■(

Reason: Essential for maintaining diverse data representations across different modalities for accurate analysis.

Q6

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

Reason: Enhances result accuracy by streamlining the sequence of information retrieval steps.

Q7

Answer presented as correct: B) FID■(

Reason: An important metric for assessing image quality representation in image captioning tasks and evaluations.