# Multimodal NLP & Vision Quiz

**Instructions:** Select the best answer(s). Multi-select where noted.

**Q1.** Visual grounding refers to:

(A) Mapping textual phrases to image regions
(B) Generating image captions
(C) Translating images to text
(D) Segmenting images at pixel level

**Q2.** (Multi-select) The CLIP model uses which training paradigm?

(A) Contrastive learning between image and text
(B) Cross-entropy language modeling
(C) Self-supervised masked image modeling
(D) Reinforcement learning

**Q3.** The multimodal attention mechanism in ViLT fuses modalities by:

(A) Concatenating visual patches and word tokens into a single Transformer stream
(B) Training separate encoders with late fusion
(C) Using a CNN backbone exclusively for images
(D) Adding textual queries to frozen image features

**Q4.** In encoder–decoder image captioning, the decoder is typically:

(A) An RNN or Transformer language model generating text
(B) A CNN classification head
(C) An object detector
(D) A feature extractor only

**Q5.** (Multi-select) Challenges in multimodal QA include:

(A) Modality alignment
(B) Table normalization
(C) Reasoning across modalities
(D) High memory consumption

**Q6.** The term *implicit decomposition* in multimodal QA indicates:

(A) Splitting complex questions into sub-queries internally

(B) Using retrieval-based pipelining

(C) Fine-tuning only the final layer

(D) Enforcing symbolic reasoning graphs

**Q7.** A common metric for image captioning that measures n-gram overlap is:

(A) BLEU

(B) FID

(C) WER

(D) AUC