PassQueen
a.k.a dumpsqueen.com

NCAGenerative AI Multimodal

EMC NCA-GENM

Version Demo

Total Demo Questions: 7

Total Premium Questions: 75

Buy Premium PDF

https://passqueen.com

support@passqueen.com

QUESTION NO: 1

An audio-language model receives batches of speech clips with different durations. Shorter clips are padded to match the longest clip in the batch. What is the primary purpose of an attention mask in this situation?

A. To prevent the model from attending to padded audio positions

B. To increase the sampling rate of each audio clip

C. To remove background noise from speech recordings

D. To convert audio frames into text tokens

ANSWER: A

Explanation:

An attention mask identifies padded positions so that the model does not treat artificial padding as meaningful audio content. During attention computation, masked positions are excluded or assigned negligible attention weight. This prevents the model from learning spurious patterns based on clip length or padding values and ensures that representations are derived from actual speech frames.

The mask does not increase the sampling rate, remove environmental noise, or align audio with a separate modality. Those functions require distinct preprocessing or model-design choices.

QUESTION NO: 2

In multimodal machine learning, what does 'early fusion' refer to?

A. Integrating different modalities at the beginning of the model pipeline.

B. Ignoring certain modalities and only using one modality for analysis and prediction.

C. Training separate models for each modality and then combining their predictions.

D. Implementing the model in the early stages of development of the ML solution.

ANSWER: A

Explanation:

Integrating different modalities at the beginning of the model pipeline is correct because early fusion combines information from multiple input types before, or at the earliest stages of, the primary learning architecture. For example, a system can transform image features, text tokens, audio features, or sensor readings into compatible representations and concatenate, align, or otherwise merge them into a shared input. The downstream model then learns from that joint representation, enabling it to identify relationships across modalities from its early layers onward. This is especially useful when meaning depends on interactions between inputs, such as associating words with image regions or relating spoken audio to visual cues. The defining characteristic is the location of integration in the pipeline: modalities are combined before substantial independent model processing has occurred. Multimodal AI systems are designed to use and reason over information from more than one data modality, and fusion architectures determine how those inputs are brought together for learning. See IBM's overview of multimodal AI and Google Cloud's discussion of multimodal AI for background on combining modalities in AI systems.

QUESTION NO: 3

What advantage does multimodal learning have over unimodal learning?

A. It requires fewer data samples for learning.

B. It can capture more complex patterns and relationships in data.

C. It is more reliable than unimodal learning.

D. It is easier to collect multimodal data than unimodal data.

ANSWER: B

Explanation:

It can capture more complex patterns and relationships in data. Multimodal learning jointly uses information from two or more modalities, such as text, images, audio, video, or sensor readings. Each modality can contribute complementary evidence: an image may provide visual context that text does not state explicitly, while text can supply semantic detail that is not readily inferred from pixels alone. Learning from these combined signals enables a model to represent cross-modal associations, align related content across modalities, and make predictions using a richer view of the underlying task. For example, a system can connect a product image with its written description, or interpret spoken language using both audio and visual cues. This broader context is a core benefit of multimodal AI and supports tasks including visual question answering, image captioning, and cross-modal retrieval. The benefit depends on appropriate data quality, modality alignment, and model design, but the ability to learn complementary and interdependent relationships across data types is the fundamental advantage. See the Multimodal Foundation Models survey and the Google Machine Learning Crash Course discussion of multimodal data.

QUESTION NO: 4

A model detects events in instructional videos by combining video frames with spoken audio. Which two (2) preprocessing practices are important for preserving meaningful audio-video relationships? Pick the 2 correct responses below.

A. Synchronising audio segments and video frames using their timestamps.

B. Maintaining the temporal order of frames and audio segments.

C. Randomly shuffling frames before they are combined with audio.

D. Pairing each video with an unrelated audio recording to increase data volume.

E. Replacing every video sequence with one randomly selected still image.

ANSWER: A B

Explanation:

Audio and video must be synchronised using their timestamps so that spoken words can be related to the correct frames. Maintaining temporal order also preserves event progression, such as an instruction being spoken before, during, or after the corresponding action is shown.

Randomly shuffling frames destroys temporal correspondence. Pairing arbitrary audio with video teaches false cross-modal associations. Reducing each video to one unrelated still image can remove the temporal evidence required for event detection.

QUESTION NO: 5

An assistant must answer questions about equipment manuals that contain explanatory text, labelled diagrams, and tables. The answer must remain grounded in the retrieved manual content. Which two (2) retrieval design choices are appropriate? Pick the 2 correct responses below.

A. Preserving links between a retrieved passage and its associated diagram or table.

B. Retrieving relevant text and visual assets using modality-appropriate or aligned representations.

C. Discarding diagrams after extracting the surrounding text.

D. Using optical character recognition as a complete replacement for visual document understanding.

E. Generating a diagram description even when no source diagram was retrieved.

ANSWER: A B

Explanation:

Preserving the association between a retrieved text passage and its source diagram or table retains context that may be necessary to interpret instructions correctly. Retrieving multimodal assets using representations appropriate to their modality, or a shared representation space where supported, enables the system to identify relevant text and visual evidence for a question.

Discarding diagrams removes potentially decisive information. OCR alone can recover some text but can lose layout, spatial relationships, symbols, and other visual meaning. Creating unsupported descriptions of a diagram would weaken grounding and can introduce hallucinated content.

QUESTION NO: 6

What is the role of CLIP (Contrastive Language-Image Pretraining) in text-to-image generation?

A. CLIP is used to generate image captions from textual input.

B. CLIP is used to convert textual input into image embeddings.

C. CLIP provides a common embedding space for both the textual and image modalities.

D. CLIP is used to enhance datasets through data augmentation for text-to-image generation.

ANSWER: C

Explanation:

CLIP provides a common embedding space for both the textual and image modalities. It is trained contrastively on paired image-and-text examples so that the representation of a caption is close to the representation of its corresponding image, while representations of nonmatching pairs are separated. This shared semantic space is CLIP's key contribution: it makes textual concepts, visual concepts, attributes, styles, and relationships comparable as vectors.

In text-to-image systems, a CLIP text encoder can transform a user prompt into contextual embeddings that condition image synthesis. For example, Stable Diffusion uses a pretrained CLIP ViT-L/14 text encoder to supply text-conditioning information to its diffusion process. As the model progressively denoises a latent representation, this conditioning guides the generated image toward visual content that matches the prompt semantically. Related architectures can also use CLIP-aligned image and text embeddings as an intermediate representation between a prompt and an image decoder.

The CLIP approach was introduced by OpenAI as a way to learn transferable visual representations from natural-language supervision; its aligned embedding space is what makes it especially useful for connecting language prompts to visual generation. See the CLIP paper and the Stable Diffusion text-to-image pipeline documentation.

QUESTION NO: 7

A team is evaluating a multimodal model that predicts a medical condition from scans and clinical notes. Which two (2) practices provide the strongest evidence that the combined model adds value beyond its individual modalities? Pick the 2 correct responses below.

A. Splitting training and evaluation data at the patient level.

B. Comparing the multimodal model with image-only and text-only baselines.

C. Reporting only one aggregate score for the combined model.

D. Measuring final performance on the data used to train the model.

E. Excluding cases for which the model produces an incorrect prediction.

ANSWER: A B

Explanation:

Patient-level data partitioning prevents records from the same patient, or closely related records, from appearing in both training and evaluation data. This reduces leakage that could make performance appear better than it would be for new patients. Comparing the multimodal model with image-only and text-only baselines is an ablation approach that measures whether the additional modality and fusion mechanism provide meaningful incremental value.

Reporting only a pooled score can conceal variation in performance and does not isolate the contribution of each modality. Evaluating on training data measures memorisation rather than generalisation. Removing difficult cases after prediction would bias the result.