

PRACTICE QUESTIONS

NVIDIA

NCA - GENM

NVIDIA-Certified Associate: Generative AI Multimodal

SHARED BY [u/Fuzzy-Kernel-1115](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	NVIDIA NCA-GENM, NVIDIA-Certified Associate: Generative AI Multimodal

Shared by [u/Fuzzy-Kernel-1115](#) on aiexamcerts

10 practice questions for the NVIDIA NCA-GENM exam, NVIDIA-Certified Associate: Generative AI Multimodal, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/nvidia/nca-genm. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCA-GENM and the question number.

1. Which component in a typical vision-language model maps visual features into the embedding space that the language model can consume?
 - A The sampler that sets the decoding temperature and the top-p cutoff for generated text tokens
 - B The tokenizer's byte-pair merge table
 - C A projector or adapter layer between the vision encoder and the language model
 - D The optimizer's momentum buffer

Answer: C

A projector, often an MLP or a small transformer, converts vision encoder outputs into token-like embeddings that the LLM accepts. The tokenizer works on text, the sampler controls decoding, and the optimizer updates weights during training, so each is a plausible but incorrect component.

2. Increasing the classifier-free guidance scale always improves image quality without any downside.
 - A True
 - B False

Answer: B (False)

Higher guidance can increase prompt adherence but may reduce diversity and introduce oversaturated or artifact-prone results. The claim of no downside is the tempting oversimplification, so the statement is false.

3. What is the main purpose of the variational autoencoder used in latent diffusion pipelines?
 - A To generate captions for training images
 - B To encode images into a smaller latent space where diffusion runs and decode results back to pixels
 - C To assign safety labels to each prompt
 - D To replace the text encoder entirely so that prompts are no longer needed for conditioning at any step

Answer: B

The VAE compresses images into latents so the diffusion process is cheaper, then decodes the denoised latent back into pixel space. Caption generation and safety labeling are separate tasks, and the VAE does not replace the text encoder that provides conditioning.

4. A learned component that maps visual features into the language model's embedding space is called a _____.

Answer: projector (also accepted: adapter, projection layer, connector)

The projector aligns vision encoder outputs with the LLM's token embedding space. The tokenizer is a tempting neighbor, but it handles text rather than visual features.

5. Which two components are typically part of a vision-language model? (Choose two.)

- A A vision encoder that turns images into feature embeddings
- B A projector that aligns visual features with the language model
- C A GAN discriminator that scores every output
- D A SQL query optimizer
- E A reinforcement learning environment simulator

Answer: A, B

Vision-language models pair a vision encoder with a projector that aligns features with the LLM's embedding space. A GAN discriminator, a SQL query optimizer and a reinforcement learning simulator are not standard parts of that architecture.

6. Which training data format is most typical for teaching a vision-language model to align images with language?

- A Only random pixel noise with fixed labels
- B Only tabular spreadsheets with numeric columns and no images or captions at all
- C Only audio recordings with no transcripts
- D Image-caption or interleaved image-text pairs

Answer: D

Image-caption pairs and interleaved image-text documents teach the model how visual content maps to language. Tabular data with no images, audio without transcripts and random noise would not teach visual-language alignment, which makes them clearly incorrect.

7. Which intermediate representation is most often produced before a neural vocoder converts speech to a waveform?

- A A binary hash of the speaker's name
- B A one-hot vector of the speaker language
- C A mel-spectrogram
- D A bounding box around the speaker's face

Answer: C

Neural text-to-speech systems often predict mel-spectrograms, and a vocoder such as HiFi-GAN turns them into waveforms. A chess board state and a face bounding box are unrelated representations, and a hash of a name carries no acoustic detail.

8. A higher mean opinion score indicates lower perceived speech quality.
- A True
 - B False

Answer: B (False)

MOS is a rating scale where higher scores mean better perceived quality. The statement reverses that meaning, so it is false.

9. A serving team wants to increase throughput of a model that receives many concurrent requests. Which technique improves GPU utilization by grouping requests together?
- A Dynamic batching
 - B Disabling the scheduler
 - C Single-request serial processing only
 - D Removing all caching

Answer: A

Dynamic batching groups compatible requests so the GPU processes them together, which raises utilization and throughput. Disabling the scheduler and serial processing reduce utilization, and removing caching usually increases compute per request.

10. Which Triton Inference Server capability helps serve several models or several versions of one model from the same server?
- A A single hard-coded model file compiled into the binary
 - B A model repository with versioned model directories
 - C A requirement that only one client can connect at a time
 - D Automatic deletion of older models after each request

Answer: B

Triton loads models from a repository organized by model name and version, so several models can be hosted together. A single hard-coded file and single-client limits are not Triton design features, and deleting models after each request would break serving.