CS109 Lecture 24: LLM Learning Guide

Diffusion

Autumn 2026

Lecture 24: LLM Learning Guide

Before working through this guide: read the Lecture 24 diffusion slides. Then open your favorite LLM and work through the concepts below **in order**. Each concept has a *Learn* prompt and a *Test me* prompt. Attempt the questions yourself before asking the LLM to grade you.

How to get the most out of this:

- Tell the LLM: "I'm a student in Stanford's CS109. We have studied random variables, conditional probability, Gaussian distributions, MLE, logistic regression, and neural networks."
- Ask it to define every symbol before using a formula.
- Explain your reasoning before asking for an answer: "Here is my reasoning: _____. Where exactly am I wrong, if anywhere?"
- Keep the focus on probability and the lecture's noise-prediction formulation. You do not need to implement a production image generator.

Watch for these confusions. Predicting noise during training is different from sampling an image during generation. The program knows the training target because it sampled the noise; the network does not receive that target as an input. Text vectors, image latents, and pixel arrays have different roles. CLIP does not draw images, and attention does not guarantee correct anatomy or relationships.

Concept 1: Images as Vectors and Adding Noise

Learn:

> "Explain an RGB image as a vector of pixel-channel values. Define every symbol in $x_t = a_t x_0 + b_t \epsilon$, with independent standard Normal entries in ϵ. Explain what the schedule controls, why x_t denotes a whole noisy image, and why noisy coordinates can lie outside the clean image's display range. Explain the difference between approaching a common noise distribution and becoming the same noise array. For independent centered unit-variance signal and noise, show why $a_t^2 + b_t^2 = 1$ keeps the variance at 1; distinguish that statement from a claim about the empirical variance of every individual image."

Test me:

> "Give me a tiny image vector, a sampled noise vector, and weights such as 0.8 and 0.6. Ask me to compute the noisy image, check the squared-weight constraint, and identify which values are random and which are chosen. Then ask me what happens at low and high noise. Wait for my answers before explaining."

Concept 2: Creating Our Own Supervised Training Targets

Learn:

> "Walk through one denoising training exercise: choose an image and caption, sample a noise level and noise array, construct the noisy image, predict the sampled noise, compute squared error, and update the weights. Clearly separate the network's inputs from its target. Explain why we know the answer key, how fresh noise produces many exercises from one image, and why the output should allow negative numbers rather than being constrained by a sigmoid. Connect gradient descent on a loss to the gradient-based training we used for neural networks."

Test me:

> "Give me a two-coordinate corruption example and a proposed noise prediction. Have me list the inputs and target, compute the sum of squared errors, and explain which quantities change during training. Include one incorrect explanation that leaks the target into the inputs, and ask me to fix it. Do not reveal the correction until I try."

Concept 3: An Estimate Versus a Sample

Learn:

> "Use a toy distribution with only two equally plausible clean images, A=[0,1] and B=[1,0]. Compute the expected squared reconstruction error for predicting A and for predicting their mean. Explain why the mean can be the best estimate while not being a possible clean image. Then explain what it means to sample from this conditional distribution. Distinguish the toy clean-image predictor from our actual noise predictor, and explain why a denoising network's estimate is not the full image sampler. Avoid claiming that blindly iterating an average guarantees a valid sample."

Test me:

> "Give me a fresh two-outcome example and ask me to find the conditional mean, compare expected squared errors, and describe how to draw a sample. Then quiz me on whether a sample must minimize reconstruction error and whether a mean must itself be a possible outcome. Have me justify each answer."

Concept 4: Generation and Random Seeds

Learn:

> "Explain generation with a trained diffusion model: start from random noise, use the network's prediction and a sampling rule, and repeat at decreasing noise levels while keeping learned weights fixed. Some samplers add randomness during updates and others do not; explain why different initial noise can still give different images. Interpret $p_\theta(\text{image} \mid \text{text})$ as a learned conditional distribution. Explain what a seed controls and which settings must be fixed for a meaningful seed comparison. Distinguish the known sampled noise in a training exercise from the absence of an exact noise answer key during generation."

Test me:

> "Ask me to sort several actions into training or generation. Then describe two different frog images generated with one prompt and ask me what could account for the variation. Ask whether weights must have changed and whether every sampling step is guaranteed to improve prompt agreement. Wait for my reasoning before grading."

Concept 5: Text, CLIP, Attention, and a Coherent Scene

Learn:

> "Use 'a frog walking on stilts across a sandy desert' throughout. Explain what each phrase asks the image to express, including relationships and pose. Explain how CLIP learns associations between images and text, why a matching image and caption can have similar vectors without identical vectors or one coordinate per concept, and how a text encoder provides contextual token representations. Give an intuitive description of cross-attention as a common way for image features to draw on relevant text information, with soft overlapping weights rather than fixed word-to-region assignments. Explain how text conditions repeated denoising predictions so that the whole scene can fit the caption. Separate the roles of text encoder, denoiser, and sampler, and note that architectures vary."

Test me:

> "Ask me why a picture containing a frog, unattached stilts, and sand does not fully satisfy the prompt. Have me explain attention in two sentences, correct the claim that CLIP draws the image, and distinguish text representations from image pixels. Then change one phrase in the prompt and ask what should change in the scene and what could remain open."

Concept 6: New Combinations, Latents, and Failure Modes

Learn:

> "Explain how learned visual patterns can generalize to a cicada eating a BLT on a catamaran in the desert even without that exact complete training scene. Explain why an undefined invented noun does not communicate a clear intended object, and why a plausible-looking scene can still have wrong counts, relationships, or anatomy. Use the whippet with an extra leg holding a tiller as an example, and explain video's additional need for consistency across frames. Finally distinguish the compressed image latent being denoised from the text representation conditioning it, and explain the final image decoder."

Test me:

> "Give me four short claims mixing up generalization, invented words, physical consistency, text vectors, and image latents. Ask me to identify and correct the mistakes. Then ask me how I would evaluate whether a generated image satisfies a multi-part prompt beyond merely containing the requested objects."

Wrap-up Prompt

> "Give me one multi-part CS109 problem that combines a small numerical noise example, identifying the network's inputs and target, computing squared loss, comparing a conditional mean with a sample, explaining seed variation with fixed weights, and using attention to connect a multi-part caption to a coherent image. Define all notation, wait for my attempt at each part, then grade my reasoning and tell me the one concept I should review most. Keep the core problem mostly conceptual. Offer the fixed-variance Gaussian likelihood interpretation of squared error only as an optional extension, and do not equate it with the full generated-image likelihood."

Final self-check: Can you explain how a program that creates noisy images with known targets can train a network whose predictions help generate new images, and how the caption influences that process repeatedly?