Last Name: _____ **First Name:** _____

SUNet ID: _____@stanford.edu

Midterm Solutions – May 1, 2026

Closed-book exam. No notes, no laptop, no phone, no internet, no AI assistance, no anything. Just your brain and a pen!

Duration: 1 hour 30 minutes **Total number of points: 100**

Instructions

Questions are grouped into logical sections to ease your thought process. There are two types of questions:

- **Multiple-choice**: One correct choice per question. Please (circle) the correct answer.
- **Free-form**: Short and concise answers.

In all cases, there is no penalty for wrong answers.

I. Diffusion with DDPM (25 points)

1. (1 point) The core unconditioned generation problem is to:
   - A. Assign each input image to one of several semantic categories.
   - B. Store every image in a compact lossless representation.
   - C. Predict a caption from an image without generating pixels.
   - D. Sample realistic images from an underlying data distribution.

2. (1 point) In DDPM, the *forward* process is:
   - A. A fixed Gaussian noising process chosen before training.
   - B. A learned U-Net that maps noise back to clean images.
   - C. A text-conditioned decoder that maps prompts to pixels.
   - D. A classifier trained to recognize clean image categories.

3. (1 point) In DDPM, the learned reverse process is usually trained to predict:
   - A. The semantic label associated with each clean training image.
   - B. The exact pixel histogram of the full training dataset.
   - C. The Gaussian noise added at the sampled time step.
   - D. The next word in a caption paired with the image.

4. (2 points) The role of the noise schedule in the DDPM forward process is to:
   - A. Set how much Gaussian noise is added at each step.
   - B. Choose the number of feature channels in the denoising network.
   - C. Convert RGB pixels into a sequence of text-like tokens.
   - D. Remove all stochasticity from the training procedure.

5. (2 points) A useful property of the DDPM forward process is that x_t can be sampled directly from x_0 because:
   - A. The reverse model is deterministic at every denoising step.
   - B. The model stores every intermediate corrupted image.
   - C. The training loss does not depend on the time step.
   - D. Sums of independent Gaussian variables stay Gaussian.

6. (2 points) The ELBO-based DDPM derivation is important because it:
   - A. Shows that image generation can be solved with nearest-neighbor lookup.
   - B. Removes all KL-divergence terms from the model.
   - C. Proves that no neural network is needed.
   - D. Connects likelihood training to denoising regression.

7. (2 points) At inference time, a DDPM sample is generated by:
   - A. Starting from a clean training image and adding noise once.
   - B. Starting from noise and denoising once.
   - C. Starting from noise and denoising step by step.
   - D. Starting from a clean training image and adding noise step by step.

8. (1 point) DDIM accelerates sampling primarily by:
   - A. Training a larger classifier on noisy images.
   - B. Skipping selected steps with deterministic updates.
   - C. Skipping selected steps with stochastic updates.
   - D. Changing the training objective to cross-entropy over labels.

9. (3+4 points) **Forward and reverse diffusion.** (i) Write the high-level mathematical form of the DDPM forward corruption process and (ii) explain why the reverse process has to be learned.

> (i) A standard formulation is $q(x_t \mid x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$, with a chosen schedule $\{\beta_t\}$. Equivalently, x_t can be sampled directly as $x_t = \sqrt{\overline{\alpha}_t}\, x_0 + \sqrt{1-\overline{\alpha}_t}\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$.
>
> (ii) The forward process is known because we choose it, but the true reverse conditional distribution depends on the unknown data distribution. The model learns a denoising rule, often by predicting the added noise, so that repeated reverse steps map noise back toward the image distribution.

10. (3+3 points) **DDPM versus DDIM.** (i) What inference bottleneck does DDIM address? (ii) What trade-off appears when taking fewer sampling steps?

> (i) DDPM generation can require many sequential denoising steps, so inference is much slower than one-shot generators such as GANs or VAEs. DDIM addresses this by using a deterministic compatible sampling process that can skip many steps.
>
> (ii) Fewer steps usually improve speed but can reduce sample quality. The sampler and step schedule become hyperparameters that trade compute for fidelity.

II. Score matching and SDEs (25 points)

1. (1 point) The score function of a distribution $p(x)$ is:
   - A. $\nabla_x p(x)$.
   - B. $\nabla_\theta p(x)$.
   - C. $\nabla_x \log p(x)$.
   - D. $\nabla_\theta \log p(x)$.

2. (1 point) Intuitively, the score function is useful for generation because it:
   - A. Directly gives the class label for a noisy image sample.
   - B. Points locally toward regions of higher log-density.
   - C. Eliminates the need to initialize from random noise.
   - D. Converts an image representation into a text embedding.

3. (1 point) Langevin sampling updates a sample by combining:
   - A. A score step plus injected Gaussian noise.
   - B. A nearest-neighbor lookup plus a class label.
   - C. A cross-attention block plus a VAE decoder.
   - D. A deterministic step plus a sampling temperature.

4. (2 points) Denoising score matching trains a model by:
   - A. Denoising samples corrupted by known Gaussian noise.
   - B. Classifying clean images into semantic categories.
   - C. Maximizing similarity between noisy images and their clean counterparts.
   - D. Solving an ODE during every training example to compute likelihood exactly.

5. (2 points) Noise Conditional Score Networks (NCSN) extend vanilla denoising score matching by:
   - A. Using only the clean-data score with no noise.
   - B. Replacing the score with a class probability.
   - C. Learning score models across many noise levels.
   - D. Removing the stochastic term from Langevin dynamics.

6. (2 points) The DDPM noise predictor and a score model are closely related because:
   - A. Both models require class labels for every training image.
   - B. The DDPM model predicts captions from noisy images.
   - C. Noise prediction determines the score up to a known scale.
   - D. Score models cannot be represented by neural networks.

7. (2 points) The reverse SDE in score-based generative modeling:
   - A. Maps data to noise without using a learned model.
   - B. Is the same thing as a supervised ViT classifier.
   - C. Requires the probability density to be available in closed form at every point.
   - D. Runs the learned score backward from noise to data.

8. (1 point) The probability flow ODE (PF-ODE) is useful because it:
   - A. Has different time marginals from the corresponding SDE by design.
   - B. Gives a deterministic sampler with the same marginals.
   - C. Removes the learned score from the model.
   - D. Forces every sampling path to be stochastic.

9. (3+4 points) **Score estimation.** (i) Define what a score model is trying to approximate. (ii) Explain why adding noise to data makes score matching practical.

> (i) A score model $s_\theta(x, t)$ approximates $\nabla_x \log p_t(x)$, the gradient of the log-density of the noised data distribution at time/noise level t.
>
> (ii) After adding Gaussian noise, the conditional corruption distribution has an analytic score, which leads to a tractable denoising objective and lets the model learn useful scores at several noise levels.

10. (3+3 points) **Reverse SDE versus PF-ODE.** (i) Compare the two sampling viewpoints. (ii) Name one practical reason to care about the PF-ODE formulation.

> (i) The reverse SDE samples with a stochastic update that uses both a score-driven drift and injected noise. The PF-ODE removes the stochastic term and follows a deterministic trajectory while preserving the same marginal distributions over time.
>
> (ii) Once written as an ODE, sampling can use numerical ODE solvers such as DPM-Solver, often requiring far fewer model evaluations than a basic step-by-step stochastic sampler.

III. Flow matching (25 points)

1. (1 point) Flow matching trains a generative model to predict:
   - A. A class label for each generated image sample.
   - B. The similarity between noisy images and their clean counterparts.
   - C. A space-dependent acceleration for speed.
   - D. A time-dependent velocity for transport.

2. (2 points) The inference goal of flow matching is:
   - A. Solve the ODE $\frac{dx_t}{dt} = u_t(x_t)$ from noise to data.
   - B. Add independent Gaussian noise until each image becomes pure noise.
   - C. Train a classifier to recognize noisy image categories.
   - D. Apply a decoder separately to every image patch.

3. (2 points) A key distinction between a velocity and a score function is that:
   - A. The velocity is always the gradient of a probability density.
   - B. Velocity moves samples through time; score climbs density.
   - C. The score is only used for supervised classification.
   - D. The velocity cannot be learned with neural networks.

4. (2 points) In a flow model, training and inference are best summarized as:

A. Train a velocity model; infer by sampling from noise and solving an ODE.

B. Train a velocity model; infer by sampling from the training data and solving an ODE.

C. Train a classifier; infer by choosing the most likely image class.

D. Train a decoder; infer by reconstructing training inputs.

5. (2 points) Conditional Flow Matching (CFM) is useful because it:

A. Requires the exact marginal velocity at every point.

B. Removes the need to sample times during training.

C. Forces all paths to be stochastic SDE trajectories.

D. Uses tractable conditional targets instead of marginal ones.

6. (2 points) For a simple straight-line interpolation $x_t = (1-t)x_0 + tx_1$, the conditional velocity is:

A. The displacement $x_1 - x_0$.

B. The score $\nabla_x \log p_t(x)$.

C. The current state x_t.

D. The one-hot class label.

7. (1 point) The motivation behind rectified flow is to:

A. Make trajectories more curved so they require more solver steps.

B. Make transport paths as straight as possible.

C. Make transport paths as long as possible.

D. Train a separate noisy classifier.

8. (3+4 points) **Conditional flow matching.** (i) Describe the training recipe at a high level. (ii) What makes this training recipe efficient?

> (i) Sample a source point x_0 from a simple prior, a target/data point x_1, and a time t. Build an interpolation x_t between the endpoints, compute the conditional target velocity, and train $u_\theta(x_t, t)$ with a regression loss to match that velocity.
>
> (ii) The model is trained by local supervised regression on sampled points and times. It does not need to simulate the full ODE during every training example or compute exact likelihoods through the flow.

9. (3+3 points) **Diffusion, score matching, and flow matching.** (i) What do these paradigms have in common? (ii) What is the main object learned by flow matching compared to score matching?

> (i) All three describe a path from a simple/noisy distribution to the data distribution and train a neural network that tells the sampler how to move along that path. They differ in the mathematical object used to define the motion.
>
> (ii) Score matching learns $\nabla_x \log p_t(x)$, which points toward higher density at a time t. Flow matching learns a vector field $u_t(x)$, which directly specifies the time derivative used to transport samples through an ODE.

IV. Multimodal guided generation (25 points)

1. (1 point) A main reason to move from pixel space to latent space is that pixel space is:
   - A. Low-dimensional and naturally aligned with semantic edits.
   - B. High-dimensional and geometrically nonsemantic for image edits.
   - C. Not aligned with text embeddings from the start.
   - D. Impossible to represent with neural network encoders.

2. (1 point) A VAE learns a latent representation by balancing:
   - A. Reconstruction and latent regularization.
   - B. Speed and accuracy of the encoder and decoder.
   - C. Quality and quantity of the latent representation.
   - D. Size and latency of the encoder and decoder.

3. (1 point) If the KL regularization term in a VAE is weighted too strongly, a common failure mode is:
   - A. Faster sampling with perfect image fidelity.
   - B. Sharper low-level texture reconstruction.
   - C. Posterior collapse of the latent representation.

D. Exact likelihood training with no approximation.

4. (2 points) Perceptual loss is used in refined image autoencoders because it:
   A. Compares only raw pixels and ignores human-visible structure.
   B. Replaces the need for an encoder.
   C. Guarantees that no artifacts can appear.
   D. Matches learned visual features instead of raw pixels.

5. (2 points) In latent diffusion, the generation model is trained:
   A. Directly on raw pixels with no encoder or decoder.
   B. In a frozen VAE latent space before decoding.
   C. Only on class labels.
   D. By predicting the next token of a text-only corpus.

6. (2 points) A Vision Transformer (ViT) adapts Transformers to images by:
   A. Treating the whole image as one scalar input.
   B. Embedding image patches as Transformer tokens.
   C. Replacing self-attention with a VAE decoder.
   D. Using only a noisy classifier during inference.

7. (2 points) CLIP-style contrastive learning aligns image and text representations by:
   A. Training only on labels from a fixed visual taxonomy.
   B. Reconstructing every image pixel from its caption.
   C. Pulling matches together and mismatches apart.
   D. Sampling with a reverse SDE and no conditioning.

8. (1 point) Classifier-free guidance differs from classifier-based guidance because it:
   A. Requires a separately trained classifier on noisy images.
   B. Works only for unconditional generation.
   C. Removes the need for a conditioning signal.
   D. Combines conditional and unconditional predictions.

9. (4+5 points) **Latent diffusion pipeline.** (i) Explain the role of the VAE encoder and decoder. (ii) Why is latent-space generation attractive compared to pixel-space generation?

> (i) The VAE encoder maps an image into a compressed latent representation used for training the generative model. At inference, after the generative model produces a latent sample, the VAE decoder maps that latent back to pixel space and supplies many low-level visual details.
>
> (ii) Latent space is lower-dimensional and more structured than pixel space, so diffusion, score matching, or flow matching can be cheaper and easier to train/sample. The decoder then handles the final conversion from compact representation to image.

10. (2+2 points) **Guidance.** (i) Compare classifier-based guidance and classifier-free guidance. (ii) State one practical limitation or cost of each.

> (i) Classifier-based guidance uses an external classifier trained on noised inputs to steer an otherwise unconditioned generative model toward a condition. Classifier-free guidance trains one model that can run both conditionally and unconditionally, then combines the two predictions to amplify the condition signal.
>
> (ii) Classifier-based guidance requires labeled data, a classifier that works on noisy inputs, and an extra classifier pass during sampling. Classifier-free guidance avoids a separate classifier but commonly requires both conditional and unconditional model evaluations per step, so sampling is still more expensive than a single unconditional pass.

*
* *

We hope you enjoyed this exam and the class so far. Looking forward to spending the rest of the quarter together!