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

SUNet ID: _____@stanford.edu

Final Exam – June 8, 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. Image generation architectures (25 points)

1. (1 point) A good image-generation backbone should be able to:
   A. Prioritize local texture while ignoring long-range structure.
   B. Use only global interactions while discarding fine spatial details.
   C. Accept conditions only after the latent has already been finalized.
   D. Capture global structure, preserve local details, use conditions, and scale.
2. (1 point) In a U-Net, skip connections are especially useful because they:
   A. Preserve local information from the downsampling path.
   B. Increase patch-token count before the attention blocks.
   C. Remove the need for any upsampling operation.
   D. Collapse all spatial features into one bottleneck vector.

3. (1 point) In adaLN-style conditioning, suppose a token representation is updated as

$$x \leftarrow x + \alpha \star \text{Operation}\big(\text{LN}(x) \star (1+\gamma) + \beta\big).$$

Which statement matches the roles of α, γ, and β?

A. α shifts $\text{LN}(x)$, γ gates the residual update, and β changes patch size.

B. α changes the sequence length, γ shifts the token, and β chooses the condition.

C. α gates the update, γ scales the normalized token, and β shifts it.

D. α removes the residual path, γ performs pooling, and β unpatchifies tokens.

4. (2 points) A Diffusion Transformer first adapts an image or latent by:

A. Treating the entire latent as one scalar token.

B. Dividing it into patches and embedding those patches as tokens.

C. Applying convolutional downsampling without tokenization.

D. Projecting only condition embeddings and ignoring image patches.

5. (2 points) Adaptive layer normalization in DiT is used to:

A. Modulate token features with quantities derived from conditions.

B. Change patch size to control the number of latent tokens.

C. Provide absolute coordinates as extra image channels.

D. Concatenate U-Net skip features in the decoder path.

6. (2 points) In DiT scaling, using a smaller patch size usually:

A. Decreases the token count while keeping the same latent size.

B. Changes only the embedding width, not the sequence length.

C. Increases FLOPs because there are more tokens.

D. Removes the need to encode positions in the attention stack.

7. (2 points) The motivation for joint attention in MM-DiT is to:

A. Prevent text and image representations from ever interacting.

B. Replace conditioning with a single global class label.

C. Keep each modality in isolated streams to prevent information leakage.

D. Let image and condition tokens interact more flexibly.

8. (1 point) Rotary position embeddings are used to:

A. Encode absolute position information as input.

B. Encode relative position information inside attention.

C. Gate normalized token features with timestep embeddings.

D. Copy high-resolution U-Net features into decoder layers.

9. (3+4 points) **U-Net versus DiT.** (i) State one architectural advantage of U-Nets for image generation. (ii) State one reason Transformers became attractive for modern image generation.

(i)

(ii)

10. (3+3 points) **Conditioning in DiT and MM-DiT.** (i) Explain the role of adaLN-style modulation. (ii) Why might joint attention be preferred for richer text/image conditioning?

(i)

(ii)

II. Model training (25 points)

1. (1 point) A modern image-generation training lifecycle is best summarized as:
   - A. Pre-train, distill, then post-training and tuning.
   - B. Tune first on a few subjects, then pre-train broadly.
   - C. Pre-train, post-train, optionally tune, and distill.
   - D. Distill first, then pre-train, post-train, and tune.

2. (1 point) The motivation for changing the timestep sampling distribution during training is to:
   - A. Emphasize hard timesteps during training.
   - B. Ensure that only nearly clean examples appear during training.
   - C. Ensure that all timesteps are sampled equally often.
   - D. Emphasize easy timesteps during training.

3. (1 point) REPA is an example of adding a loss that:
   A. Changes timestep sampling to oversample easy examples.
   B. Adds preference rewards directly to final samples.
   C. Distills many teacher steps into one student step.
   D. Aligns model features with a pretrained image representation.

4. (2 points) Curriculum learning for image generation typically moves from:
   A. Long prompts and high resolution to short prompts and low resolution.
   B. Easier settings to harder settings as training progresses.
   C. Randomly mixed difficulty to a staged easy-only dataset.
   D. Fixed resolution and prompt complexity throughout training.

5. (2 points) Preference tuning methods such as ReFL, Flow-GRPO, and Diffusion-DPO aim to:
   A. Use human or reward-model preferences to steer generation.
   B. Use the image generation model to tune human preferences.
   C. Prefer a single subject identity via the use of a rare token.
   D. Compress a many-step sampler into a preferred, smaller-step student model.

6. (2 points) DreamBooth-style tuning is mainly used to:
   A. Improve general quality through broad continued training.
   B. Train every parameter from scratch on a large generic dataset.
   C. Bind a few subject images to a rare token.
   D. Speed up sampling by matching teacher transitions.

7. (2 points) LoRA is useful for tuning because it:
   A. Requires all original model weights to be retrained fully for maximum quality.
   B. Adds full-rank updates to every layer in the backbone of the model.
   C. Changes prompt wording at inference time without the need for explicit retraining.
   D. Trains small low-rank adapters instead of all weights.

8. (1 point) In image generation, distillation most often tries to:
   A. Increase the number of sampling steps for higher fidelity.
   B. Produce good samples with fewer sampling steps.
   C. Specialize a model to one specific subtask.
   D. Train a smaller student model on the larger teacher's outputs, using the same number of sampling steps.

9. (3+4 points) **Training lifecycle.** (i) Name the four broad stages of a modern image-generation training lifecycle. (ii) Give the high-level purpose of each stage.

(i)

(ii)

10. (3+3 points) **Distillation.** (i) Why is teacher/student distillation attractive for diffusion or flow models? (ii) Name one limitation or trade-off of an approach such as progressive distillation, consistency models, DMD, or adversarial distillation.

(i)

(ii)

III. Evaluation (25 points)

1. (1 point) A generated image can be visually realistic while ignoring the prompt, or follow the prompt while looking poor. This motivates evaluating:
   A. Diversity and safety.
   B. Speed and memory use.
   C. Aesthetics and prompt adherence.
   D. Text rendering and editing fidelity.

2. (1 point) Pairwise human comparisons are often easier than absolute ratings because they:
   A. Ask raters which of two images is better.
   B. Ask raters to explicitly calibrate a five-point scale that can then be used to rank all images.
   C. Require each image to match one reference exactly.
   D. Produce single-sided evaluations that are easy for diagnostics.

3. (1 point) Elo-style ratings improve on raw win rate by:
   A. Averaging binary pass rates through an easy-to-compute formula.
   B. Counting only wins against the same fixed opponent.
   C. Treating every opponent as equally informative.
   D. Accounting for the strength of the opponent.

4. (2 points) FID evaluates generated images by:
   A. Comparing raw pixel distances to a reference image.
   B. Comparing real and generated feature distributions.
   C. Asking a prompt-adherence judge yes/no questions.
   D. Averaging human pairwise preference ratings.

5. (2 points) A CLIPScore-style metric may give similar scores for an image paired with the prompts "A cute teddy bear is reading a book" and "A cute book is reading a teddy bear," even though the relation is reversed. What limitation does this illustrate?
   A. Reference-free metrics require a ground-truth image before they can be computed.
   B. Feature-distribution metrics cannot estimate whether images look realistic.
   C. Global image-text alignment can miss compositional or relational swaps.
   D. Pairwise human ratings cannot compare two images generated from related prompts.

6. (2 points) Suppose an edited image preserves the right object and style, but the object is shifted a few pixels relative to the reference. Which evaluation concern applies?
   A. Pixel-wise metrics such as MSE can over-penalize small spatial shifts.
   B. Pairwise human ratings cannot compare two images from the same prompt.
   C. Reference-free metrics require the shifted reference image.
   D. Elo-style ratings become exact only when the shift is small enough.

7. (2 points) A best practice for MLLM-as-a-Judge evaluation is to:
   A. Avoid tuning it with human preferences to avoid introducing bias.
   B. Keep A/B ordering fixed across pairwise comparisons for determinism.
   C. Ask for the final score before the evidence.
   D. Ask for the evidence before the final score.

8. (1 point) A prompt says, "A cute teddy bear is reading a book." Suppose an MLLM-as-a-Judge first asks whether there is a teddy bear, whether there is a book, and whether the teddy bear is reading the book. This decomposition is useful because it:

   A. Uses the same grading rubric across prompts.

   B. Makes it easier to localize failures of specific claims.

   C. Ensures that the final score is unbiased.

   D. Removes the need to inspect the generated image.

9. (2+2+3 points) **Human evaluation.** (i) What signal does a pointwise rating record? (ii) What signal does a pairwise rating record? (iii) Why can Elo-style updates be more informative than raw win rate?

   (i)

   (ii)

   (iii)

10. (3+3 points) **Metrics and judges.** (i) Why are fixed metrics such as FID, CLIPScore, SSIM, or LPIPS useful but incomplete? (ii) What is the role of MLLM-as-a-Judge methods?

    (i)

    (ii)

IV. Trending topics (25 points)

1. (1 point) Many strong open-weight text-to-image generators share the following high-level design pattern:

   A. Pixel-space U-Nets with no latent bottleneck or text encoder.

   B. VAE latents with convolutional blocks and text embeddings trained from scratch.

   C. Flow training with VAE latents but no multimodal conditioning.

   D. Flow training, MM-DiT, VAE latents, and strong text embeddings.

2. (1 point) Suppose an open-weight text-to-image model performs close to leading closed models on standard image-generation evaluations. Why is this significant?
   A. It proves open-weight models are safer than closed models.
   B. Strong generation capabilities become easier to inspect, adapt, and deploy outside a few labs.
   C. It means evaluation is no longer needed once model weights are public.
   D. It shows that video and image generation now have identical difficulty.

3. (1 point) A core added difficulty in moving from image generation to video generation is:
   A. Handling time, temporal consistency, and larger compute.
   B. We now need a time tokenizer to encode the time dimension.
   C. The lack of spatial structure caused by multiple frames.
   D. Evaluating one static frame with image-only metrics.

4. (2 points) Fréchet Video Distance is used as:
   A. A spatio-temporal VAE architecture choice for compression.
   B. A method for enforcing frame consistency during sampling.
   C. A video evaluation metric analogous in spirit to FID.
   D. A prompt format for describing camera motion and duration.

5. (2 points) A guardrailed image-editing pipeline can avoid full regeneration by:
   A. Regenerating the full image from the edited prompt.
   B. Ranking fully regenerated candidates by aesthetic score.
   C. Ignoring the source image and generating from text only.
   D. Using a VLM to choose actions for editing software.

6. (2 points) For diffusion language models, a practical analog of image noise is:
   A. Gaussian noise added to discrete token identities.
   B. Masking a subset of discrete tokens.
   C. Dropping punctuation while keeping all semantic tokens.
   D. Shuffling all tokens while treating order as unchanged.

7. (2 points) Diffusion LLM inference is **best** described as:
   A. Starting masked, then predicting, committing, and remasking tokens.
   B. Predicting tokens one by one from left to right.
   C. Filling every mask in one pass and discarding low-confidence tokens from the final sequence.
   D. Applying Gaussian noise to token representations and decoding once at the end.

8. (1 point) In fill-in-the-middle code completion, a model sees a prefix and suffix and fills the missing span. Why is this natural for diffusion-style language models?
   A. A left-to-right model can use the suffix without any special setup.
   B. Each missing token can be filled independently in one pass.
   C. The missing span can be masked, then predicted and revised using both sides as context.
   D. Training uses next-token likelihood only, so masking is unnecessary.

9. (3+4 points) **Images, videos, and editing.** (i) What changes when extending image generation to video generation? (ii) Why might an editing-action pipeline be safer than fully regenerating an edited image?

   (i)

   (ii)

10. (3+3 points) **Diffusion for text.** (i) Contrast autoregressive LLM generation with diffusion-style LLM generation. (ii) Name one advantage and one challenge of diffusion LLMs.

    (i)

    (ii)

*
* *

We hope you enjoyed spending this quarter with us. Happy graduation if this was your last quarter. Wishing you a relaxing summer and the best of success ahead!