Research Proposal: **Failure Modes of Efficient Diffusion Models**

Student Name: **Yue Li, Hengyi Li**
Course/Section: **STAD68**
Date: **2026-02-09**

Submission Requirements (Mandatory). You must submit **both**:

- A compiled PDF of this proposal (`.pdf`)
- The LaTeX source used to generate it (`.tex`) **plus any required supporting files** (e.g., `.bib`, figures, style files) so the document compiles.

Submissions missing either the `.pdf` or the `.tex` (and required supporting files) are considered **incomplete**.

---

1 Topic

- **Problem statement:** Recent advances in diffusion-based image generation increasingly prioritize efficiency, enabling high-quality samples with dramatically fewer inference steps. Techniques such as consistency models and aggressive sampling acceleration have made diffusion models practical for real-time and large-scale deployment. However, these efficiency-oriented approximations fundamentally alter the behavior of diffusion models and may introduce systematic failure modes that are poorly understood.

  Unlike catastrophic failures, such issues often manifest subtly: reduced diversity, overconfident generation of canonical outputs, brittle behavior under compositional prompts, or failures that evade standard evaluation metrics. This project investigates the failure modes of efficient diffusion models, aiming to understand how and why such failures arise under efficiency constraints.

- **Motivation:** As efficient diffusion models are increasingly deployed in real-world systems, understanding their limitations becomes as important as improving their performance. Existing evaluations predominantly focus on perceptual quality or text–image alignment, which may obscure deeper structural failures. A systematic analysis of failure modes can inform both evaluation practices and future model design, particularly in settings where diversity, robustness, and reliability are critical.

- **Scope:** This project adopts a failure-mode–driven perspective. We aim to define and analyze a taxonomy of failure modes induced by efficiency constraints in diffusion models. While the taxonomy is broad, empirical evaluation is selective and diagnostic rather than exhaustive. The project focuses empirically on inference-time acceleration methods, with Latent Consistency Models (LCMs) serving as a representative case study.

  Training-time efficiency methods, architectural compression, and non-image generative models are outside the scope of this work.

Optional (recommended): We first propose a conceptual taxonomy of efficiency-induced failure modes in diffusion models. We then empirically analyze one representative failure mode in depth and provide qualitative or illustrative evidence for others, using modern diffusion pipelines as controlled testbeds.

2 Summary of Selected Papers

Paper 1: High-Resolution Image Synthesis with Latent Diffusion Models

Citation: [1]
Link (optional): `https://arxiv.org/abs/2112.10752`
Key points and relevance:

1 We first train an autoencoder that is much smaller in size compared with other types of diffusion models. We compress the image by autoencoder to a much smaller latent space. Then we train the model in the latent space. It takes much less time to train the model in the latent space compared to in the pixel space. After the training is complete, we decode the image from latent space to pixel space and output the image.

2 Latent diffusion models require less computational cost compared with other popular models such as Autoregressive models, GANs and pixel-based diffusion models but yet to produce high-quality image generation.

3 After the autoencoder is trained, it could be used later for other tasks and other image generations. By storing the trained antoencoder, it also saves time in the future.

It will help us to write a related work paragraph linking this paper to latent diffusion model. This paper acknowledges the trade-off but does not analyze its failure modes. This gap is exactly where our project comes in.

Paper 2: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis

Citation: [2]
Link (optional): `https://arxiv.org/abs/2307.01952`
Key points and relevance:

1 Unlike traditional stable diffusion models, SDXL model significantly improves a lot by producing images with better quality. SDXL consistently outperforms other models due to its huge neural network structure, more specifically, due to its three times larger UNet backbone and its redesigned structure. In addition, those improvements do not require additional supervision.

2 SDXL adopts more powerful text encoders, which are CLIP ViT-L and OpenCLIP ViT-bigG. These encoders have larger context dimension, allowing the model to better understand text prompts, making the model better understand user's need, and improving the quality of images.

3 Stable diffusion models require a minimum image resolution. So, many small images are dropped or forced to be upscaled. In either way, the quality of images will be compromised. SDXL model, on the other hand, conditions the model on the original image height and width during training, allowing it to learn how resolution affects image appearance.

The paper explicitly states the failure modes of SDXL model for instance, it mixes two different objects that do not match in a single image; it fills up the wrong color to a subject (say it fills up a banana with blue color); it makes mistake when generating small, detailed, or complex parts of an image; it has trouble generating long, readable, and consistent text inside images. These are interesting points to our research topic.

Paper 3: Consistency Models

Citation: [3]
Link (optional): `https://arxiv.org/abs/2303.01469`
Key points and relevance:

1 Traditional diffusion model is slow because it takes too many steps or iterations. Consistency model is proposed because it takes much less steps but still produces good quality images. It generates images in one step by directly predicting noisy images. It outperforms previous diffusion distillation methods, achieve state-of-art FID and work well on benchmarks.

2 Consistency models learn a self-consistent mapping that sends any noisy point along a diffusion ODE trajectory directly to the clean data which is the input, enabling single step image generation while still supporting multi-step refinement for improved quality

3 Experiments across CIFAR-10, ImageNet 64x64, and LSUN datasets show that consistency distillation significantly outperforms progressive distillation in few-step generation and achieves state-of-the-art FID in one and two step settings. Moreover, it achieves competitive results compared to diffusion models despite only requiring one or two forward passes.

Collapsing a full trajectory into one mapping means removing intermediate structures, removing gradual denoising refinement and forcing the model to learn a global shortcut. As a result, this can lead to mode collapse like behavior, reduced diversity, over-smoothing and canonical outputs.

Paper 4: Phased Consistency Models

Citation: [4]
Link (optional): `https://arxiv.org/pdf/2405.18407`
Key points and relevance:

1 Phased Consistency Models (PCM) identifies structural failure modes in Latent Consistency Models (LCMs), a popular approach for accelerating diffusion models to few-step generation. The paper argues that LCMs suffer from three core issues: (1) inconsistency across inference steps, where outputs vary significantly with different step counts due to stochastic trajectory modeling; (2) poor controllability, especially instability under large classifier-free guidance (CFG) and weak sensitivity to negative prompts, caused by training–inference mismatch during guided distillation; and (3) inferior low-step performance, since LCMs rely on pointwise L2-style losses without enforcing distribution-level alignment.

2 To address these limitations, PCM divides the PF-ODE trajectory into multiple sub-trajectories ("phases") and enforces consistency locally rather than globally, enabling deterministic multi-step sampling. Additionally, it introduces an adversarial consistency loss in latent space to improve few-step quality.

3 For our project on failure modes of efficient diffusion models, this paper is directly relevant because it formalizes trajectory compression and guidance mismatch as structural causes of instability in accelerated diffusion methods.

Paper 5: TLCM: Training-efficient Latent Consistency Model (2024)

Citation: [5]
Link (optional): `https://arxiv.org/html/2406.05768v5?utm_source=chatgpt.com`
Key points and relevance:

1 While latent diffusion models achieve high-quality images, their inference requires tons of computational power. Existing acceleration methods such as consistency model that we talked about earlier, either suffer from noticeable quality degration or require large-scale real data and long training time. This is why TLCM is designed to generate high quality images with minimal additional computation and data.

2 Unlike the previous models, intead of requiring real image and text pairs, TLCM samples Gaussian latent noise and uses a pretrained diffusion model (e.g., SDXL) to denoise it into synthetic latent states, which are then used as supervision to train the TLCM. This enables distillation entirely from samples generated by pretrained diffusion models without access to the original training dataset.

3 A 3-step TLCM trained without any real data achieves a higher text-image alignment (CLIP 33.68) and more aesthetic quality than existing 4-8 step accelerated diffusion models, demonstrating that data-free latent consistency distillation can preserve teacher-level generation quality under aggressive few-step sampling.

Since TLCM tries to keep image quality high while greatly reducing the number of sampling steps, it is a good example for studying how few-step latent generation changes the underlying image distribution. It also helps us examine whether common evaluation metrics such as CLIP and FID may hide deeper structural problems in the generated images. This directly supports the goal of this project, which is to identify and understand failure modes caused by efficiency-driven acceleration in diffusion models.

Paper 6: Distilling Diversity and Control in Diffusion Models (2025)

Citation: [6]
Link (optional): `https://arxiv.org/html/2503.10637v4?utm_source=chatgpt.com`
Key points and relevance:

1 Disrilled diffusion models generate images much faster but lose sample diversity because they commit to the final image structure very early in the generating process, even though they still retain the internal representations needed for diversity. To solve this, we can implement a simple hybrid method by using the base model for only the first time points before switching to the distilled model. By doing so, it restores diversity while maintaining computational efficiency and shows that diversity loss is caused by generation dynamics rather than lost knowledge.

2 It introduces a trajectory visualization method to inspect what the model believes the final clean image will be at each time point. THis method shows that diffusion models can temporarily represent some semantic elements such as a cat face or an football, and later discard

them during denoising. Distilled models commit to the final image structure almost immediately, while base models refine structure gradually across many time points. This explains why distilled models suffer diversity collapse and why the first time point becomes the main bottleneck.

This paper is directly relevant to our research because it identifies a key efficiency induced failure mode in distiled diffusion models which is sample diversity collapse. It shows that although distilled models remain fast and score well on standard metrics, they commit to image structure too early. This reduces variation across seeds. In our research, we can dive deeper into this aspects and explore more interesting facts.

3 Goal / Target Outcome (Required)

The primary goal of this project is to identify, categorize, and analyze failure modes of efficient diffusion models. Rather than treating failures as isolated artifacts, we aim to understand them as systematic consequences of efficiency-oriented approximations.

Deliverables

- **Primary deliverable(s):** (e.g., method, reproduction + extension, benchmark, theoretical result)
  - A failure mode taxonomy for efficient diffusion models
  - A detailed empirical case study of a representative failure mode
  - Qualitative and quantitative evidence illustrating additional failure modes
- **Artifacts:** (e.g., code repo, trained models, evaluation scripts, report)
  - Reproducible inference and evaluation code
  - Generated image samples
  - Diagnostic plots and visualizations
  - A final written report synthesizing conceptual and empirical findings

Success Criteria

- **Conceptual Criteria:**
  - Clear definitions and distinctions between failure modes
  - Plausible connections between efficiency assumptions and observed failures
- **Empirical Criteria:**
  - Quantitative evidence for at least one failure mode
  - Qualitative or illustrative evidence for additional modes
- **Metrics / Tools:**
  - LPIPS diversity
  - CLIP embedding variance
  - CLIP text–image alignment
  - Visual inspection of representative samples

4 Feasibility Plan (Required, Detailed)

This section must evaluate feasibility and explicitly discuss: **(1) repos/codebases, (2) datasets, (3) playgrounds/tools, (4) compute/resources.** Include concrete details and a fallback plan.

4.1 Repositories / Codebases (Required if relevant)

- **Candidate repo(s):**
  - `https://github.com/huggingface/diffusers` Hugging Face diffusers
  - Official LCM LoRA adapters for Stable Diffusion
- **What you will use it for:**
  - Stable Diffusion as a controlled baseline
  - LCM-based accelerated sampling for efficiency comparison
- **Feasibility assessment:**
  - Install/run status (what you've tried so far)
  - Documentation quality and ease of reproduction
  - Activity/maintenance, licensing constraints (if any)
  - Expected modifications needed

4.2 Datasets (Required if relevant)

- No external datasets are required.
- Evaluation is performed using manually designed prompts that probe known challenging aspects of image generation, such as diversity, compositionality, and fine-grained structure.
- All prompts are synthetic and do not involve sensitive or personal data.

4.3 Playgrounds / Tools (Required if relevant)

- PyTorch
- Hugging Face diffusers and transformers
- LPIPS for perceptual diversity
- CLIP for semantic feature analysis
- Google Colab for experimentation and visualization

4.4 Compute / Resources Needed (Required)

- **Hardware:** CPU/GPU/TPU requirements (type/count if known) Single GPU (e.g., NVIDIA T4 or A100 via Colab)
- **Estimated runtime:** (per experiment/training run; expected number of runs) seconds per prompt per configuration
- **Storage needs:** (datasets, checkpoints, logs) Around 1 GB for images and logs
- **Where compute will come from:** (local machine / lab servers / cloud / free tiers) Colab free tiers
- **Fallback plan if compute is limited:** (smaller models, fewer runs, reduced scope, alternative methods) Reduce number of prompts or samples if compute is constrained.

5 Preliminary Results (Required)

Provide **at least one** of:

- **Preliminary empirical results:** (small-scale experiments, baseline runs, sanity checks, partial replications.) **Initial experiments comparing standard diffusion sampling with accelerated LCM sampling reveal clear evidence of distributional failure under aggressive efficiency constraints.**
  - Sample diversity decreases monotonically as the number of inference steps is reduced.
  - LPIPS diversity and CLIP feature variance drop sharply for highly accelerated sampling.
  - CLIP text–image alignment remains relatively stable across configurations.
- **Preliminary theoretical results:** formal problem setup, derivations, proof sketches, complexity analysis.

  Problem Setup:
  Let P_T denote the distribution induced by T-step diffusion sampling, and P_k the distribution induced by accelerated $k << T$-step sampling.
  Let efficiency-induced failure be $\Delta_k = \mathcal{D}(P_k, P_T)$, where $\mathcal{D}$ is a function of capturing structural diversity, stability, and controllability.
  Failure Modes and Proof Sketches:
  - Diversity Collapse: If accelerated sampling is a more contractive mapping $x_0 = H_k(x_T)$, then the Jacobian norm $||J_k||$ decreases. By convariance propagation, $Cov[x_0] \approx J_k Cov[x_T] J_k^T$, so output variance shrinks $\Rightarrow$ reduced LPIPS variance and CLIP feature spread.
  - Canonical Output Bias: Distillation often minimizes L2-style loss. It favors conditional means over rare modes in a multimodal distribution. The model plays it safe. It makes normal, common outputs, but avoids unusual or rare ones.
  - Early Commitment Instability: Let t^* be commitment time. In accelerated models, $t_k^* << t_T^*$. Early commitment reduces sensitivity and correction capacity. Because of that, it makes diversity collapse very quickly.
  - Classifier Free Guidance Over Amplification: When guidance strength is large, it makes the model push too hard toward the prompt. In accelerated models, this also magnifies errors, causing unnatural or distorted images.

* Loss of Fine Details: Small details such as text, thin lines need many gradual correction steps to become clear and accurate. However, when we reduce the process to only a few steps, then model does not have enough chances to fix these small errors. As a result, find details are more likely to look blurry, distorted, or incorrect.

Complexity
Let C denote the cost of one neural network forward pass.
Let T denote total number of denoising steps in the full diffusion model.
Let k denote the number of denoising steps in the accelerated model, where $k << T$.
The complexity of full diffusion: $\mathcal{O}(TC)$.
The complexity of accelerated diffusion: $\mathcal{O}(kC)$.
Even though $k << T$, as k decreases, the sampling mapping becomes more compressed and contractive, which reduces diversity and stability.

What you have done so far

- We carefully reviewed six relevant research papers and identified strong connections to our research topic such as trajectory compression, diversity collapse, guidance instability, and distributional shift under few step sampling.
- We successfully installed and ran diffusion pipelines using Hugging Face diffusers, including both the standard diffusion model (teacher model) and the accelerated LCM-based sampling model (student model).
- In preliminary theoretical result, we showed monotonic decrease in sample diversity as inference steps are reduced, significant drop in LPIPS diversity and CLIP feature variance under aggressive acceleration and relatively stable CLIP text-image alignment despite diversity degradation.
- Based on the proposal and papers, we have begun organizing failure modes into different categories:
  - Distributional failures such as diversity collapse and canonical bias
  - Structural failures such as fine-detail degradation and compositional gragility
  - Dynamical failures such as early commitment and step inconsistency
  - Control failures such as Classifier-Free Guidance instability and weak negative prompt sensitivity

Evidence (tables/figures encouraged)

To assess the feasibility of our failure-mode–driven approach, we conducted an initial empirical study comparing standard diffusion sampling with accelerated Latent Consistency Model (LCM) sampling under varying inference step counts (30, 8, 4, and 2 steps).

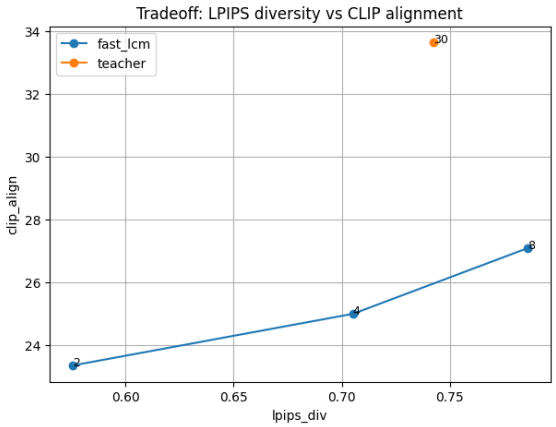
We evaluated models using two diversity metrics—LPIPS perceptual diversity and CLIP embedding variance—as well as CLIP text–image alignment for semantic fidelity.

	model	steps	prompt_id	lpips_div	clip_featvar_div	clip_align
0	teacher	30	0	0.862286	0.257986	35.286781
1	fast_lcm	2	0	0.608096	0.276579	25.049267
2	fast_lcm	4	0	0.735723	0.333315	28.417471
3	fast_lcm	8	0	0.715120	0.236375	33.727352
4	teacher	30	1	0.753750	0.168811	32.370277

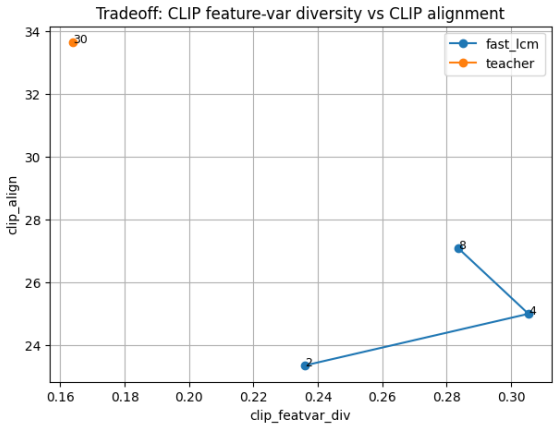
Aggregated (mean over prompts):

	model	steps	prompt_id	lpips_div	clip_featvar_div	clip_align
0	fast_lcm	2	3.5	0.575565	0.235883	23.340149
1	fast_lcm	4	3.5	0.705233	0.305436	24.987922
2	fast_lcm	8	3.5	0.785831	0.283610	27.079212
3	teacher	30	3.5	0.742426	0.163800	33.639407

Figure 1: Aggregated Diversity and Alignment Metrics Across Sampling Regimes.



(a) Diversity–Fidelity Tradeoff Under Accelerated Sampling (LPIPS-Based Diversity)



(b) Semantic Diversity vs Alignment Under Efficient Diffusion Sampling

Figure 2: Diversity–Fidelity Tradeoff

Key Observations

- Diversity decreases under aggressive acceleration. LPIPS diversity shows a clear monotonic decline as the number of sampling steps decreases. The 2-step configuration exhibits substantially lower perceptual diversity than the baseline model, indicating strong distributional collapse.
- Alignment degrades more gradually. While CLIP alignment scores decrease with fewer steps, the degradation is smoother and less severe than the drop in diversity. This suggests that accelerated models retain coarse semantic correctness even as they lose distributional coverage.
- Diversity metrics are not identical. LPIPS diversity and CLIP embedding variance exhibit partially differing behavior, reflecting the distinction between perceptual variation and semantic

dispersion. This supports our hypothesis that commonly used metrics may capture different aspects of model behavior.

- **implications:** These preliminary results provide concrete evidence of a representative failure mode: efficient diffusion sampling induces distributional collapse before catastrophic semantic misalignment occurs.

  This supports our broader project direction of analyzing failure modes not as isolated artifacts, but as systematic consequences of efficiency-oriented approximations.

6 Detailed Plan (Required)

Provide a step-by-step plan with milestones. A week-by-week plan is recommended.

1. **Milestone 1:** Establish a reproducible teacher-student pipeline and generate images under fixed configurations such a fixed number of steps in image generations. Design structured prompts and compute diversity and alignment metrics.
   Expected output: A pipelien that can be easily run again, with comparison tables and simple plots showing the trade-off between diversity and image quality.

2. **Milestone 2:** Carefully test important problems like loss of diversity, committing too early to one structure and instability when CFG is large. We will change the number of steps, random seeds and CFG values to see how the models behave under different configurations.
   Expected output: Clear numbers and side-by-side images showing at least one failure that we can explain with confidence.

3. **Milestone 3:** Improve mathematical definition of failure and explain why fewer steps can cause problems. Connect our experiments to ideas like shrinking variance, L2 loss pushing toward average outputs, early structure locking, and the tradeoff between speed and stability.
   Expected output: A clear explanation of why making the model faster for instance from $\mathcal{O}(TC)$ to $\mathcal{O}(kC)$, can reduce diversity and make outputs less stable.

4. **Milestone 4:** Group all failure types into clear categories such as distribution, structure, dynamics and control. Combine theory and experiments into one final report with clear explanations.
   Expected output: A complete failure-mode framework, clear figures and a comprehensive written report.

7 Potential Pitfalls and Mitigations (Required)

List key risks and what you will do if they occur.
Note that this list may not include all possible pitfalls. The ones below are simply those we have considered so far.

- **Pitfall 1:** It is possible that the diversity drop or instability is not strong enough to clearly show a failure.
  Mitigation: We will increase the number of prompts or samples, test more extreme step reductions or we will focus on one failure mode in deeper details rather than covering too many of them.

- **Pitfall 2:** The diffusion pipeline or the implementation of the latent consistency model may behave strangely or differently than expected. Or, the results might not be reproducible.
  Mitigation: We will simplify the setup, use examples in the Hugging Face official website, fix random seeds, and reduce the system of a minimal working baseline if it is necessary.
- **Pitfall 3:** Running many prompts and configurations may take too long on Google Colab free version.
  Mitigation: We will reduce the number of prompts or samples, use smaller image sizes, or test fewer configurations while keeping the main comparison. Alternatively, we might also consider purchase Google Colab Pro to shorten the running time.
- **Pitfall 4:** Scores like CLIP or LPIPS may stay stable even when images look worse, making it hard to quantify some failures.
  Mitigation: We will show example images and clearly point out the problems, not just rely on numbers.

8 Expected Workload (Required)

Estimate your workload and break it down by activity.

Weekly time estimate

- Total hours/week: 16-20 hours
- Per person: 8-10 hours/week

Time breakdown

- Reading / paper digestion: 3-4 hours/week
  Both members review papers and discuss theoretical ideas together.
- Implementation / engineering: 4-6 hours/week
  Hengyi will focus more on pipeline setup and evaluation scripts.
  Yue will supports testing and verification.
- Experiments / debugging: 5-6 hourse/week
  We split experimental runs such as different step settings, Classifier Free Guidance sweeps, prompt groups and we cross check results.
- Writing / documentation: 4-5 hours/week
  We devide sections of the report i.e. empirical versus theoretical parts and then revise together for consistency.

In a word, we will pretty much work together for every task and will perform cross check for each other's result.

Heaviest components

– What will take the most time and why?
  Experiments and debugging will definitely take the most time. Because running multiple configurations on limited GPU resources will be slow. They also requires repeated testing and comparisons.

– What dependencies might block progress (compute, data access, etc.)?
GPU availability on Google Colab will be the main dependency. Other things like session time limits, memory issues or long download time may also delay experiments. Furthermore, extra time may be required to understand certain models if they are not sufficiently clear to us.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021.

[2] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[3] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023.

[4] Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models, 2024.

[5] Qingsong Xie, Zhenyi Liao, Zhijie Deng, and Haonan Lu. Tlcm: Training-efficient latent consistency model for image generation with 2-8 steps. *arXiv preprint arXiv:2406.05768*, 2024.

[6] Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. *arXiv preprint arXiv:2503.10637*, 2025.