Technical Report

Improving Prompt Alignment in Text-to-Image Diffusion Models

A Hybrid Approach Combining CONFORM and Feynman-Kac Diffusion Steering

Autor: Samuel Alejandro Fernández Martínez

Email: samuelfernandezmatinez@gmail.com

1 Abstract

Recent advances in text-to-image diffusion models have achieved remarkable generative capabilities, yet many outputs still fail to align with the intended prompt, often producing malformed structures or missing entities. To address this issue, this work explores a hybrid methodology that combines two complementary approaches during the generation process. The first, **CONFORM** (*Contrast is All You Need for High-Fidelity Text-to-Image Diffusion Models*), tracks the evolution of individual entities and reshapes them when overlaps occur. The second, **Feynman-Kac Diffusion Steering**, periodically generates multiple candidates in parallel, selects the most promising sample, and iteratively refines it through successive generations. Building on recent findings that diffusion models can be understood in three distinct phases, this research integrates both methods into a unified workflow. Preliminary evaluations on multiple diffusion backbones including Stable Diffusion v1.5, Stable Diffusion XL, and Latent Consistency Models suggest that the combined approach yields qualitatively higher fidelity compared to each method in isolation. While further quantitative validation is required, the proposed fusion demonstrates practical potential for more reliable prompt adherence in generative image systems.

2 Feynman–Kac Diffusion Steering (FK steering)

Overview and core idea

Feynman–Kac Diffusion Steering (FK steering) Singhal et al., 2025, arXiv:2501.06848, is an **inference-time, particle-based** framework for steering diffusion models toward samples that maximize an arbitrary reward function. Rather than retraining the model, FK steering runs multiple interacting diffusion trajectories (particles) in parallel and periodically **resamples** a particle according to scores computed from user-defined *potentials* (intermediate rewards) and remove the ones with lower score. This resampling bias the sampling distribution toward high-reward outcomes while remaining training-free.

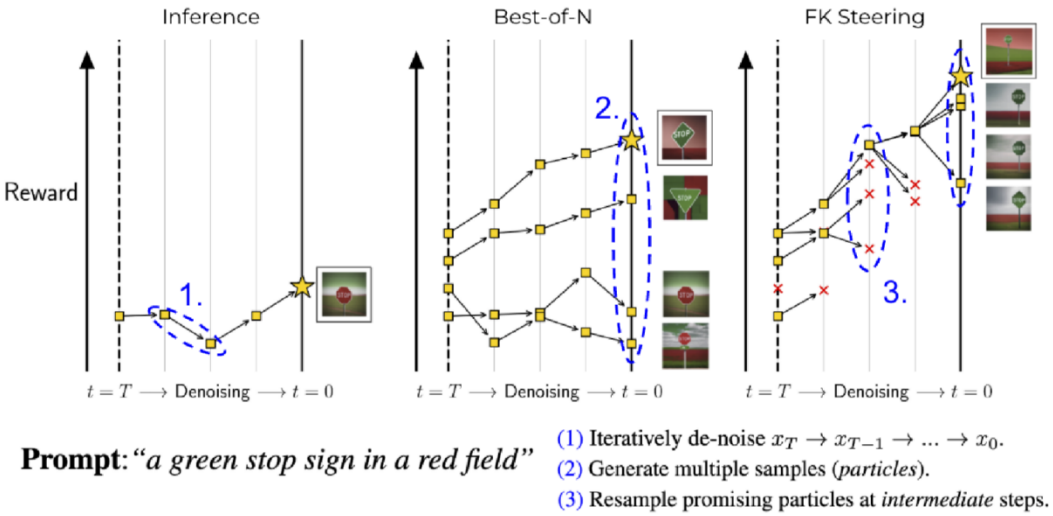

Figure 1: A General Framework for Inference-time Scaling and Steering of Diffusion Models.

Problem addressed

Standard diffusion sampling techniques often rely on either naive single trajectory sampling or simple heuristics such as best of n selection and classifier free guidance. These strategies face several limitations:

- **Inefficient search under complex objectives.** Best-of-n sampling scales poorly with objective difficulty: when the target reward is rare (e.g., unusual attribute combinations, fine-grained prompt adherence), exponentially more samples are required to find a satisfactory outcome.
- **Inability to optimize non-differentiable rewards.** Classifier based or gradient guided steering requires a differentiable signal, making it unsuitable for objectives like human preference scores, aesthetic judgments, or black-box evaluation functions.
- **Weak utilization of intermediate information.** Conventional methods typically evaluate only the final outputs, discarding potentially useful signals at intermediate denoising steps, which slows down convergence toward high-fidelity candidates.

Methodology

- **Particles and trajectories:** FK steering maintains a population of particles, each representing a candidate diffusion trajectory from noisy state to denoised sample.
- **Potentials (intermediate rewards):** At preselected timesteps the method evaluates a potential for each particle, this can be a learned reward, a CLIP based prompt fidelity score, a classifier output, or any heuristic indicating future reward likelihood.
- **Resampling:** Particles are reweighted and resampled according to these potentials; high-scoring particles are duplicated while low scoring ones are discarded, concentrating compute on promising regions of the sample space.
- **Options and design choices:** The framework supports various choices of potentials, intermediate reward definitions, and resampling/sampler strategies; these design choices trade off compute, diversity, and bias toward the reward.

Key contributions and empirical behavior

- **Training-free steering:** FK steering improves control and sample quality at inference time without fine-tuning or additional training.
- **Arbitrary rewards:** Because potentials can encode differentiable or no differentiable objectives, FK steering enables steering for human preference models, attribute control, or other bespoke rewards.
- **Practical gains:** The original study reports that FK steering can outperform larger fine-tuned models on prompt fidelity (for example, steering a smaller 0.8B model to beat a 2.6B fine-tuned model in certain prompt-fidelity tests) and that only a small number of particles is often sufficient to produce large improvements in practice.

Practical considerations and limitations

- **Compute vs. benefit:** FK steering uses more inference compute than single-sample generation (more trajectories to simulate), though effective resampling often means the method is more sample efficient than retraining or massive best of n sampling. Choosing particle count and resampling schedule is an engineering tradeoff.
- **Reward design:** The method's success depends on well chosen intermediate potentials poorly designed rewards can bias generation undesirably or reduce diversity.

3 CONFORM: *Contrast is All You Need for High-Fidelity Text-to-Image Diffusion Models*

Overview and core idea
CONFORM, Meral et al., 2023, arXiv:2312.06059, proposes a training free, test time contrastive mechanism applied to cross attention maps of text-to-image diffusion models. The core intuition is to use a contrastive objective (InfoNCE style) over attention activations so that tokens and attributes that should co-occur remain tightly associated while distinct objects become segregated in attention space. This guidance is applied during the denoising process to encourage the model to maintain consistent, disentangled attention for different prompt concepts.

Problem addressed
Diffusion generators often fail to realize complex multi-concept prompts: objects can be omitted, attributes can be misassigned, and distinct entities can collapse or overlap in the final image. CONFORM targets precisely these failure modes by operating on attention maps the intermediate mechanisms that mediate how text tokens influence image regions to reduce confusion between tokens and preserve correct object-attribute bindings. In some studies like Marioriyad et al., 2025, arXiv:2410.20972 explain how one of the main sources and biggest one for this problem is how an entity overlap another one and don't let it generate properly which CONFORM helps to solve this problem.

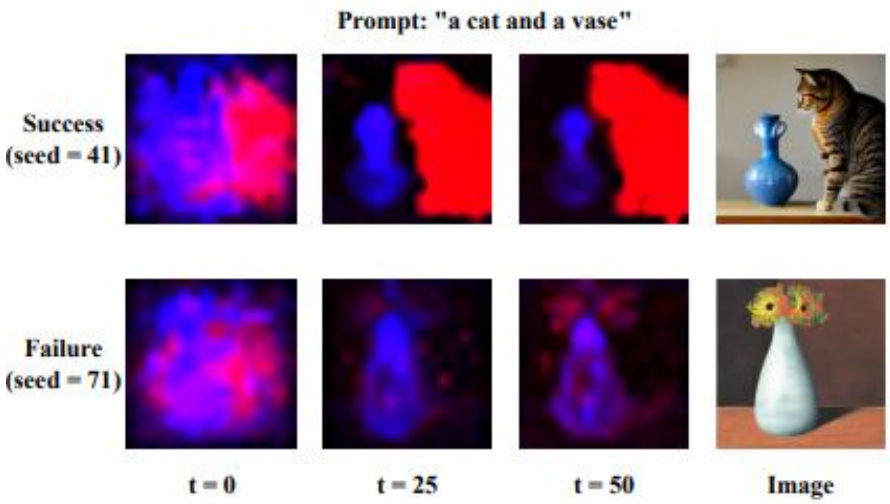

Figure 2 Attention Overlap Is Responsible for The Entity Missing Problem in Text-to-image Diffusion Models! Case where the cat entity is overlapped by the vase entity.

Methodology

- **Attention extraction:** At selected denoising timesteps, cross-attention maps linking text tokens to spatial latents are extracted.
- **Positive / negative pair construction:** Pairs of attention vectors are labeled as *positive* when they correspond to semantically related token pairs (e.g., "red" ↔ "apple") and *negative* when unrelated (e.g., "red" ↔ "car" when not intended).
- **Contrastive loss at test time:** An InfoNCE-style objective is computed on these attention vectors across nearby timesteps; the loss attracts positive pairs and repels negatives. Minimizing this objective (via a small, test-time optimization) reshapes attention patterns to be more semantically faithful.
- **Iterative application:** The contrastive adjustment is applied repeatedly across denoising steps so attention consistency is enforced throughout generation, reducing drift and entity overlap.

Key contributions and empirical behavior

- **Training-free and model-agnostic:** CONFORM improves fidelity without retraining and works with both latent (e.g., Stable Diffusion) and pixel-space models (e.g., Imagen).
- **Improved multi-concept fidelity:** The authors report consistent gains in prompt adherence across diverse multi-entity prompts and demonstrate improvements on automatic proxies for fidelity (e.g., CLIP/related scores) as well as human preference evaluations.
- **Practicality:** Because the method operates at inference time and manipulates attention (rather than model weights), it can be incorporated into existing generation pipelines with modest additional compute per sample.

Practical considerations and limitations

- **Compute overhead:** Test-time optimization on attention maps adds inference cost compared to a vanilla single pass. The method trades extra compute for higher fidelity, so selection of how many timesteps and how much optimization to run is an engineering choice.
- **Reward / pair design sensitivity:** The effectiveness depends on correctly identifying positive and negative pairs from the prompt; ambiguous prompts or complex linguistic structure may require careful tokenization or prompt parsing.
- **Not a universal fix:** CONFORM mitigates many failure modes tied to attention entanglement, but other issues (e.g., dataset biases, architectural limitations) may still cause errors that attention shaping alone cannot resolve.

4 Manifolds, Random Matrices and Spectral Gaps — *The geometric phases of generative diffusion*

Overview.
The work *Manifolds, Random Matrices and Spectral Gaps: The geometric phases of generative diffusion* Ventura et al., 2024, arXiv:2410.05898, analyses the latent geometry of diffusion models under the manifold hypothesis by studying the spectrum of the Jacobian of the score function. Using tools from random matrix theory and statistical physics, the authors show that the eigenvalue spectrum exhibits ***gaps*** whose structure reveals distinct submanifolds and that these spectral signatures naturally partition the generative process into three qualitative phases. These phases explain how diffusion models separately form manifold geometry and internal density, helping to account for why diffusion avoids certain forms of manifold overfitting.

The three generative phases (summary).

1. **Trivial phase (Phase I)**: Early in the reverse diffusion trajectory noise dominates and the model behavior is effectively trivial: perturbations are dominated by isotropic components and no clear manifold structure is yet formed. The spectrum at these timesteps shows no separation indicative of low dimensional structure.
2. **Manifold coverage phase (Phase II)**: As denoising proceeds, the dynamics begin to capture the manifold-internal distribution: the model fits local data variability along manifold tangents. Spectral analysis reveals emerging structure (sub-gaps) that correspond to tangent directions and internal variability; the model shapes the internal density without yet fully projecting onto the data support.
3. **Consolidation phase (Phase III)**: At later timesteps the score function becomes largely orthogonal to the manifold: dynamics consolidate samples onto the data support and project particles off noisy directions onto the learned manifold. The spectrum shows clear gaps that separate manifold tangent modes from orthogonal directions. This is the stage where final correctness and support level properties are enforced.

Why these phases suggest a split application of CONFORM and FK steering.
The two methods you fuse attack fidelity from complementary angles:

- **CONFORM** shapes and stabilizes *attention-based* token region associations during denoising by applying a contrastive objective on cross-attention maps at inference time. Its mechanism directly targets entanglement between tokens and attributes and enforces consistent attention over timesteps (i.e., it reduces confusion as entities form). Because this effect is most useful while entities and their local structure are being formed i.e., during the phase when the model is building manifold-internal structure it is natural to apply CONFORM early and through the middle of generation. Moreover, applying CONFORM from the very beginning helps the attention dynamics evolve in a disentangled way so later phases have better-formed candidates to refine.

- **Feynman–Kac (FK) steering** is a particle-based, resampling strategy that steers sampling by periodically selecting high-scoring trajectories (particles) using intermediate potentials. FK steering is especially powerful when the sampling dynamics can benefit from competitive selection that is, once particles are close to or on the data manifold and small differences among them determine final fidelity. That description matches Phase III (consolidation), where projection onto the support and final cleaning occur. In this phase, resampling high-quality particles yields a strong improvement in final prompt adherence because the candidates differ in fine-grained fidelity rather than gross structure.

Putting those observations together, a principled scheduling is:

Practical scheduling heuristic (recommended):

1. **Apply CONFORM starting from the earliest denoising steps** (Phase I) and continue through the beginning and bulk of Phase II. The goal is to keep attention associations disentangled as the model moves from noise toward manifold coverage. Early application helps prevent malformed entity emergence and attribute swap errors that are hard to correct later.

2. **Switch to FK steering once the process nears consolidation** (later part of Phase II → Phase III). Use FK steering to run multiple particles in parallel and perform periodic resampling so the most promising samples are concentrated and refined. This capitalizes on FK steering's strength in selecting the best final candidate(s) for fidelity.

How to detect the switching point in practice.

- **Spectral-gap monitor (preferred, principled):** compute a lightweight estimate of the Jacobian/score-spectrum or a proxy (e.g., singular values of a score-Jacobian approximation or a statistic derived from attention/Jacobian correlates). The geometric

paper shows that **opening spectral gaps** signals the transition toward manifold coverage and later consolidation; use the relative size or first clear drop in the sorted singular values as a trigger to transition from CONFORM→FK. This is the most theoretically grounded heuristic because it directly uses the signal the paper identifies. However, calculating this in the inference means an increase of computational cost.

- **Practical, simpler heuristics if Jacobian estimates are too costly:** switch based on either (a) a fixed denoising-step threshold calibrated per model/dataset (e.g., after X% of steps), or (b) when a prompt-fidelity proxy (CLIP similarity or other internal score) stops improving under CONFORM and plateaus — then begin FK steering to exploit particle selection. These heuristics are coarser but often adequate in applied settings.

Caveats and trade-offs.

- CONFORM adds test time optimization on attention maps extra compute early on while FK steering multiplies inference work by running particles in parallel. The fusion trades compute for fidelity; choose particle counts and CONFORM optimization budget to balance throughput and quality.
- The spectral-gap strategy is principled but requires implementing Jacobian/spectrum probes; practical proxies may be necessary for production pipelines.

Short Conclusion.
The random-matrix analysis and phase decomposition introduced by *Manifolds, Random Matrices and Spectral Gaps* provides a principled explanation for *when* different inference interventions are likely to be most effective. Guided by that theory, using CONFORM from Phase I into early Phase II helps form disentangled, well-attended entities, while switching to FK steering in the consolidation regime lets particle selection amplify final prompt fidelity. This principled scheduling both matches the methods' mechanisms and is supported by the spectral-phase diagnostics proposed by the paper.

5 Overall fusion strategy

One practical limitation of Feynman–Kac (FK) diffusion steering is that it relies on repeated generation and resampling until a high-reward sample is found. In constrained settings for example, when the allowed number of denoising steps or wall clock time is limited FK steering may fail to produce a satisfactory image because none of the sampled latents ever develops correctly separated and well formed entities, it's even explained in the manifold study mentioned before that as the phases advance, the changes get more limited. In other words, no amount of resampling will succeed because randomness of the resampling may not find good enough sample and each step makes it harder to get big changes.

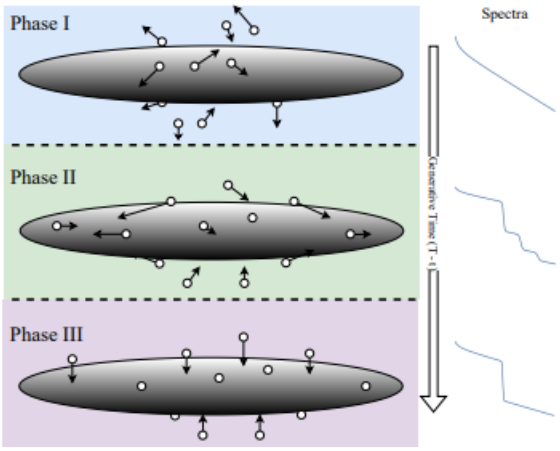

Figure 3 Manifolds, Random Matrices and Spectral Gaps: The geometric phases of generative diffusion. The three different phases and the particles behavior in every phase.

By contrast, CONFORM operates on attention dynamics and is often effective at disentangling distinct entities and preserving correct object attribute bindings even in early phases. However, CONFORM can sometimes produce images that, while structurally more faithful to the prompt, are visually less coherent or less stylistically attractive compared to standard outputs. Each method therefore addresses fidelity from a different angle: CONFORM focuses on structural separation via attention shaping, while FK steering focuses on selective refinement via particle competition.

The fusion proposed in this project leverages these complementary strengths. Concretely, we apply CONFORM during early to mid denoising steps to encourage disentangled attention patterns and to increase the probability that candidate latents contain correctly separated entities. Once candidates have reached a state where fine-grained differences determine final fidelity (i.e., near the consolidation phase), we switch to FK steering: multiple particles are spawned from the CONFORM-improved latents, scored by a prompt-fidelity potential, and resampled iteratively to concentrate compute on the most promising trajectories. This two-stage pipeline reduces the number of resampling rounds required by FK steering and tends to produce final images with higher prompt adherence than either method used alone. Its also possible to apply both at the same time and it would get even bigger quality but it will make it slower and more costly but it's an option.

Expected benefits of this fusion include:

- Faster convergence to a perceptually acceptable image (fewer FK resampling cycles required).
- Greater final prompt fidelity by attacking the problem from complementary directions (attention disentanglement + particle selection).

- Improved robustness in low-compute or low-step budgets, since CONFORM increases the chance that a small set of particles already contains well-formed candidates.

Trade-offs and practical considerations:

- The fusion increases inference cost relative to a single vanilla pass: CONFORM introduces test-time attention optimization and FK steering multiplies particle trajectories. Tuning the CONFORM budget and particle count is therefore necessary to balance throughput and quality. As well as making the process of generation slower.
- The switching point between CONFORM and FK steering is a hyperparameter. Principled triggers include spectral-gap or Jacobian-based diagnostics; simpler alternatives are fixed step thresholds or a plateau in a prompt-fidelity proxy (e.g., CLIP score).
- The effectiveness of the fusion depends on good prompt parsing and correct positive/negative pair design for CONFORM; poorly designed pairs can limit gains or introduce biases.

This fusion is implemented following the phase decomposition described in *Manifolds, Random Matrices and Spectral Gaps: The geometric phases of generative diffusion*. Empirically, CONFORM produces the largest gains during the earliest denoising steps: it quickly improves token region disentanglement and reduces entity overlap while the model is still forming manifold-internal structure. As denoising continues, the marginal benefit of additional CONFORM iterations tends to diminish. Conversely, Feynman–Kac (FK) steering is most effective once candidate latents are already reasonably formed typically in the later part of Phase II and during Phase III because FK relies on evaluating and selecting among multiple plausible images.

For a more stable and effective fusion, therefore, we concentrate CONFORM activity in Phase I and the early portion of Phase II to shape attention and encourage correctly separated entities. Once the generation enters the consolidation regime (late Phase II → Phase III), control is handed off to FK steering: multiple particles are spawned from the CONFORM improved latents and iteratively resampled according to a prompt-fidelity potential. This handoff leverages CONFORM's structural corrections to increase the likelihood that FK steering's particle population already contains high quality candidates, reducing the number of resampling cycles required and improving final prompt adherence.

In practice, the switch can be determined by a principled spectral gap or Jacobian based diagnostic, or by practical proxies such as a fixed denoising-step threshold or a plateau in a prompt-fidelity metric (e.g., CLIP similarity). This phase-aware scheduling balances stability and compute: CONFORM reduces structural failure modes early, and FK steering concentrates refinement effort where it is most impactful.

6 Latent bootstrapping via large-model base generation

Concept.
The fusion above highlights how strongly the generation outcome depends on the *initial latents* the pipeline works from. Building on that insight, we propose a latent bootstrapping strategy that leverages a larger, high capacity model to produce well formed base latents, then uses a smaller model together with the CONFORM→FK steering pipeline to refine and improve the final image while keeping inference cost manageable.

Workflow intuitive:

1. Use a large, high-capacity model to generate a base image. This model produces latents that already contain strong, well-separated structure but may be expensive to run repeatedly to fix certain problem with the image such as wrong entity.
2. Use an inversion procedure with a smaller model to map the base image back toward noisy latents (an approximate reverse-diffusion / inversion). The goal is to obtain a starting latent that the small model can continue from.
3. From the inverted latent, perform a guided denoising pass with the fused CONFORM → FK steering pipeline on the smaller model: apply CONFORM during the early/mid denoising steps to disentangle attention and then switch to FK steering for final particle based refinement.
4. Optionally iterate this process (e.g., regenerate base with different seeds or conditioning) if additional exploration is desired.

Why this helps.

- A large model is used only once (or a small number of times) to produce *high-quality* seed latents, which increases the likelihood that the small-model particles already contain correct entity separation among other qualities for the pic.
- The smaller model then performs the compute-cheaper inversion and a fusion denoising, which reduces overall compute compared with using the fusion on the large model while often producing a final image that is *better than the original base image*.

Figure 4 Before is SD XL, after is using the new fusion with model dreamshaper which is smaller model. It solved the dragon being stuck in the castle and the roof being red.

- This approach opens the door to applying the same fusion idea to other image manipulations to go even further improvements (style transfer, targeted edits, cross-model refinement) because the small model plus inference-time controls can be used to re-shape or re-steer the seeded latents.

More interactions with FK steering, Conform and initial latents.

It is important to emphasize the role of properly configuring **CONFORM's hyperparameters** and introducing **random noise into the initial latents**. Without these adjustments, the combined method can fail to deliver meaningful improvements.

Case 1: FK Steering only, starting from step 0 or 1 (no CONFORM applied).
The generated result is clearly flawed in this case, the intended dragon is not visible but just a red horse.

a photo of a brown horse and a red dragon

Case 2: CONFORM + FK Steering, but without initial randomization for every particle.
The result improves, but every trajectory converges to the same image because all particles started from the exact same initial latent. As a result, FK Steering becomes ineffective and

computationally wasteful.

Case 3: Adding randomized noising to the initial latent for each particle, combined with CONFORM and FK Steering.
This approach allows the system to explore a wider range of possibilities, producing more diverse and higher-quality generations.

Lastly, its recommended to limit the use of Conform for late latents starts such as 70 to 100. The reason can be seen in this from the research papers, ***figure 3***, if we look at the image as a cloud of dusts, composed by different particles, we can see in the early phases it can change and adapt with the most freedom possible as they don't have yet a direction to go while later it has bigger difficulty. Conform makes big aggressive changes to where the model wanted to generate the different entities, while in early phases the model can adapt to those changes, later it just lead to a distortion in the image and incapability of generating the entities in the new zone nor in the original zone.

It would be wrong to think it is bad to use Conform in later phases but only if it has been using since the start, ergo from the first step. The reason is simply in those cases, the model and Conform objective agree in a big percentage where each entity should stay, this means only little changes will happens in the images made by Conform and like the ***figure 3*** shows, it can still adapt the model to changes if they are really small at late phase. However, if you used Conform in early phases, doing later may have no big impact enough to compensate the computation and time that it cost.

Potentially, if we cycle this process with both Conform, FK steering and random noising, we could get better images if we repeat it in different generations, it could be something such as:

- Do one generation and choose the best image.

- Start again with the same image as initial laten and random noising for the different particles.
- If the final image of the new run is better than the last winner, try again with said imagen.
- To avoid getting stuck in the same set images, it's possible to perform different hyperparameters to explore new zones, perhaps with more InfoNCE, looking for no just overlapping but a percent of enough attention for certain entity.

This large process in multiple generation has no been tested yet as it would require too many computing and time that surpass the limit of a more common use of this image generators but it's possible.

Acceleration with fast samplers.
Models and samplers that support *few-step* generation (for example, latent-consistency-style samplers or other accelerated diffusion schedules) can make the base generation and the whole bootstrap pipeline much faster. Using a latent consistency model to create the base image reduces the cost of the initial large-model pass and enables more aggressive CONFORM/FK hyperparameters on the small model because the total step budget remains practical. Note, however, that FK steering still benefits from a modest number of timesteps to evaluate and resample particles reliably; extremely aggressive step reduction can degrade FK performance.

Latent Consistency Model does indeed work with Conform or FK steering, it's only necessary to use this mechanism until the image is good enough and then the model can proceed to track the final imagen fully denoised. For easy images, you can do this in just a few steps.

Figure 5 Replicating the results in the research of FK steering with just a few steps thanks to latent Consistency Models and Conform

Practical considerations & caveats

- **Inversion fidelity:** The quality of the inversion (how well the small model maps the base image back to a useful latent) determines how much of the large-model structure is preserved. Use a deterministic or near-deterministic inversion (DDIM/DDRM style) adapted to the small model to maximize retention.
- **When to start:** It's useful to don't start the denoising from the image obtained by the large model from a completely noised image, while you still get advantages by doing so, it's also a good benefit to start by a late time step and complete the rest of the steps with the fusion FK steering and conform to solve little problems such as the before pic ***Figure 4***, for example start at step 70 or 80 out of 100 total steps.
- **Compute tradeoffs:** The approach reduces repeated large-model runs but adds inversion and extra small-model inference. It is beneficial when the large model is substantially more expensive than the small one.
- **Hyperparameter tuning:** Key knobs are (a) how many large-model base generations to run, (b) the inversion fidelity/steps, (c) CONFORM optimization budget, and (d) FK steering particle count and resampling schedule. Tuning these jointly is necessary to obtain the best cost/quality tradeoff.
- **Suitability:** This pipeline is attractive for research/production settings where a single high-quality seed can be afforded and where many variations/refinements are desired at lower marginal cost.

Conclusion:
Latent bootstrapping leverages the strengths of large and small models together: the large model supplies high-quality structural latents, while the small model plus CONFORM→FK steering provides cost-efficient, controlled refinement. When combined with fast samplers (e.g., latent-consistency methods) and careful inversion, this pipeline offers a practical path to higher-fidelity outputs without running expensive particle-based steering at large model scale for every sample and keeping a relative fast generation.

7 Demo

The following pics were made all with the same model under the same procedure and hyperparameters using the fusion of FK steering and Conform in stable diffusion v1.5 100 steps. Being the last one showing no just the best particle but the other options for the final step, 4 particles in this case.

A photo of a white lamp and a yellow lamb

A photo of a pink book and a green airplane

a photo of a brown horse and a red dragon

a photo of a brown horse and a blue dog

a photo of a brown horse and a blue dog