MIDTERM UPDATE

Task & Motivation & “Aether”

Latent Space Transport Control

Nov 2025

Task Statement

Develop a reinforcement learning framework to **steer diffusion & flow-matching models** away from unsafe concepts during inference, without retraining the base model.

Motivation – The Urgency

- **The Context:** Recent proliferation of non-consensual deepfake platforms and AI-generated violence has exposed the failure of text-based filters.
- **Cost Efficiency:** Supervised Fine-Tuning (SFT) is computationally expensive and degrades general capabilities.
- **The Focus:** Our benchmark specifically targets high-harm concepts: Violence, Gore, and Nudity.

What is Aether?

Aether is a Framework: We are building a reinforcement learning framework that trains a specialized Steering Agent (Policy π_ϕ).

The Name: In physics, "Aether" was the hypothesized invisible medium for light. Here, it represents the Latent Space: the invisible, fluid medium our agent steers through to shape generation.

Related Work

- **Song et al. (2021):** Score-Based SDEs. Established generation as a Probability Flow ODE.
- **Lamba et al. (2025):** RL for Alignment. Proposed safety rewards for diffusion.
- **Wagenmaker et al. (2025):** Latent Space RL. Demonstrated steering policies for robotics/diffusion.

Models, Tools & Novelty

Methodology

Tools & Stack

The Base: A frozen Diffusion/Flow-Matching model treated as a deterministic *Probability Flow ODE (Ordinary Differential Equation).*

Methodology

We intervene in the deterministic flow.

- **Concept Detection:** Linear Probing (Alain & Bengio) to find safety boundary.
- **Steering Agent:** PPO (Proximal Policy Optimization) $\pi(z,t)$ that outputs transport vectors.
- **Reward:** Optimal Transport (W2 cost) + Safety Penalty.

Novelty

Layer Sensitivity Analysis

We don't steer blindly. We use Linear Probing to**"Semantic Single Boundary" (SSB)**.

- **Optimal Transport Reward:** Minimizes the energy cost to ensure we don’t break the image structure.

r

Analysis & Evaluation

Benchmarks

Benchmark Setup

Evaluation is performed using a custom dataset of adversarial prompts.

- **Dataset:**
  The Unsafe Benchmark (I2P):
    Target: Violence, Gore, Nudity.
  The Control Group (MS-COCO): For Utility Preservation.
- **Baseline:** Unsteered Flow Model
- **Target:** Steered Model

Evaluation Metrics

We measure success across 4 axes:

SSR	LPIPS	W2	FPR
SAFETY RATE (↑)	QUALITY LOSS (↓)	TRANSPORT COST (↓)	OVER CORRECTION (↓)

Hypothesis

By optimizing for Transport Cost (W_2), our model will achieve high Safety (SSR, Safety Success Rate) with significantly lower perceptual distortion (LPIPS, Learned Perceptual Image Patch Similarity) compared to standard guidance methods, by lowering the Wasserstein distance (W2), while maintaining near-zero intervention on safe data (Low FPR, False Positive Rate).