Exploratory Study: Alternative Control Strategies

Sriram Bharadwaj
MT2024114

December 15, 2025

1 Overview and Motivation

While the primary project methodology focuses on Nishad Bagade's experiments, I conducted an extensive parallel study to evaluate the limitations of different control signal modalities. This section documents my progression through three distinct control paradigms:

1. **Structural Control:** Attempting to disentangle geometry from texture in vehicle respraying.
2. **Discrete Semantic Control:** Using bounding box layouts to guide composition.
3. **Continuous Color Control:** Using abstract spatial maps ("Mosaic") to guide style and atmosphere.

This comparative study highlights critical challenges I encountered in ControlNet training, specifically regarding "Concept Bleeding" and "Spatial-Semantic Mismatch."

2 Study I: Structural Disentanglement (Vehicle Respray)

My initial objective was to develop a *Virtual Vehicle Respray* system capable of modifying vehicle paint while preserving chassis geometry using Canny Edge ControlNet.

2.1 Engineering Stabilization: The Manual Loop

Initial attempts using high-level abstractions (HuggingFace `accelerate`) caused persistent "Mixed Dtype" crashes due to conflicts between AdamW (Float32) and model weights (Float16). I transitioned to a manual training loop using raw PyTorch with `GradScaler`:

- **Frozen Components:** VAE, U-Net, and Text Encoder cast to Float16.
- **Trainable Components:** ControlNet maintained in Float32 for gradient stability.

This architecture enabled stable training on consumer GPUs (Tesla T4) without OOM errors.

2.2 The "Ghost Car" Phenomenon & Structural Forcing

Early inference yielded a "Generic Sedan" failure mode, where the base model's prior overpowered the control signal (e.g., a Lamborghini input became a generic white Mercedes). To counter this, I introduced **Structural Forcing** via prompt dropout:

$$P(\text{prompt} = \text{""}) = \lambda_{structure} \approx 0.7 \tag{1}$$

By removing text conditioning during training, I forced the model to rely solely on edge features to resolve the latent image. This resulted in the "Ghost Car"—a gray, geometrically perfect reconstruction of the input, confirming structural locking.

2.3 Failure Analysis: The "Gray Labeling" Trap

In the final validation, Red Ferraris were consistently reconstructed as Silver sedans. Root cause analysis traced this to the data preprocessing: my K-Means color extractor misclassified metallic paints (high specularity) as "Gray." The model effectively learned the mapping `Image: Red Ferrari` $\rightarrow$ `Text: "Gray Car"`, reinforcing the dataset bias. I mitigated this during inference using a "Force-Feeding" strategy with high guidance scales (> 8.5) and prompt weighting.

3 Study II: Discrete Layout Control (Semantic Bounding Boxes)

Following the structural experiments, I investigated **Discrete Semantic Control**. I implemented a ControlNet conditioned on a canvas of colored bounding boxes corresponding to COCO object categories, aiming to control scene composition explicitly.

3.1 Methodology Refinements

To improve data quality, I engineered several enhancements to the layout generation pipeline:

- **The Painter's Algorithm:** I sorted annotations by area, drawing the smallest objects last to prevent occlusion by larger bounding boxes.
- **Golden Angle Coloring:** I replaced random RGB assignment with deterministic Golden Angle hue generation to maximize visual separability between classes.
- **Instance Borders:** Black separation borders were added to prevent the merging of adjacent instances (e.g., two people becoming one blob).

3.2 Failure Analysis: Semantic Bleeding

Despite these improvements, the Layout model revealed a critical theoretical limitation: **Concept Bleeding**.

- **The Problem:** When a small object box (e.g., "Cat") was placed inside a larger box (e.g., "Person"), the model frequently merged the concepts. Instead of a person holding a cat, the model generated a person wearing a cat-patterned shirt.
- **Prior Dominance:** The Stable Diffusion prior often overruled the spatial layout. If a "Frisbee" box was placed near a "Person," the model would often hallucinate a "Dog" nearby, regardless of the layout, due to strong dataset correlations.

4 Study III: Continuous Color Guidance (The Mosaic Approach)

To resolve the issues of Concept Bleeding found in discrete layouts, I pivoted to **Continuous Color Control**. I hypothesized that abstract color maps could guide composition without triggering semantic conflicts.

4.1 Attempt A: Global Palette Conditioning (Failure)

Initially, I conditioned the model on horizontal color strips representing the global palette.

- **The Sepia Trap:** The model exhibited severe mode collapse, converging to a muddy, vintage aesthetic.
- **Root Cause:** This was a **Spatial-Semantic Mismatch**. ControlNet is a spatially-aligned architecture (pixel-to-pixel). Feeding it abstract, non-spatial strips forced the model to learn an impossible mapping, causing it to revert to the average mean color of the dataset (brown/sepia).

4.2 Attempt B: Spatial Color Map (Success)

I corrected the spatial mismatch by developing the "Mosaic" signal: downscaling ground truth images to a grid (e.g., 32×32) and upscaling them back. This preserved spatial correspondence while abstracting detail.

- **Hardware Optimization:** To train on the WikiArt dataset with limited VRAM (15GB), I utilized **Gradient Accumulation** (simulating batch size 4) and **Gradient Checkpointing** on the ControlNet module.
- **Results:** Unlike the Layout approach, the Mosaic signal is *class-agnostic*. A red block in the lower left could become a red sofa, a fire, or a sunset depending entirely on the prompt. This successfully eliminated semantic bleeding and allowed for robust "Composition Transfer" (e.g., applying the lighting composition of a street scene to a living room generation).

5 Comparative Conclusion

This study demonstrates a clear trade-off between discrete and continuous control signals.

1. **Discrete Controls (Layouts)** suffer from semantic conflict; when the control signal (Box: Cat) conflicts with the base model's prior (Context: Person), the model tends to merge features destructively.
2. **Continuous Controls (Mosaic/Canny)** provide superior fidelity because they decouple semantics from structure. The Mosaic approach proved most effective for artistic workflows, as it constrains *where* things are (via color distribution) without enforcing *what* they are, leaving the semantic interpretation to the text prompt.