# From Erasure to Transplant: Probing Embedding Semantics via Semantic Surgery

**Giulia Pietrangeli** (*2057291*) - **Lorenzo Musso** (*2049518*)
Advanced Machine Learning and Computer Vision (2025)
*Based on the framework by Xiong et al. (NeurIPS 2025)*

Academic Year 2025/2026

**Abstract**

Text-to-Image (T2I) diffusion models have demonstrated remarkable capabilities in generating high-quality images, yet controlling specific semantic attributes without retraining remains a challenge. Building upon the *Semantic Surgery* framework [Xiong et al., 2025], originally designed for concept erasure, this work proposes a paradigm shift towards **Semantic Transplantation**. We introduce a novel vector injection mechanism that allows for the surgical replacement of concepts within the latent embedding space. Additionally, we address the non-linearity of hyperparameter optimization by introducing a comparative study between Classical Machine Learning and Deep Learning approaches to automate the prediction of surgical parameters $(\lambda, \alpha)$. Our experimental analysis validates the method's efficacy while exposing intrinsic model biases, offering a diagnostic probe into the latent rigidity of Stable Diffusion.

## 1 Introduction

Precise semantic control in Text-to-Image models, such as **Stable Diffusion** [Rombach et al., 2022], typically requires fine-tuning or optimization. We extend the training-free *Semantic Surgery* framework [Xiong et al., 2025], originally designed for erasure, to perform **Semantic Transplantation**. By injecting target concept vectors, we enable surgical replacements directly in the CLIP embedding space [Radford et al., 2021]. Our contributions include:

1. **Method Validation:** Assessing structural integrity in object and context swapping.
2. **Ablation & Stress-Testing:** Mapping the $(\lambda, \alpha)$ landscape and quantifying latent biases.
3. **Automation:** Comparing ML and DL approaches to predict optimal surgical parameters.

## 2 Methodology: From Erasure to Transplant

We extend the original Semantic Surgery framework from erasure to **Vector Injection**. Operating on CLIP text embeddings $e_{in} \in \mathbb{R}^{L \times D}$, we introduce a target concept $e_{new}$ to surgically replace a source $e_{src}$. The transplant operation is defined as:

$$e^* = e_{in} + \lambda \cdot \mathbf{M}_\alpha \odot (e_{new} - e_{src}) \tag{1}$$

where $\lambda$ scales the injection force, $\odot$ denotes element-wise multiplication, and $\mathbf{M}_\alpha$ is a spatial mask derived from token-wise similarity, thresholded by sensitivity $\alpha$. This formulation enables precise local editing while preserving global context, unlike standard prompt engineering.

## 3 Investigation I: Method Validation

We validated the method through Subject and Context Swapping, analyzing structural and semantic integrity.

### 3.1 Subject Swap: The Shape Bias

Experiments revealed a critical dependency on the geometric prior of the source object. Morphologically similar swaps (e.g., *Dog* $\to$ *Cat*) achieved high spatial consistency (IoU $\approx 0.88$). However, structural divergences exposed a latent **Shape Bias**: the *Bear* $\rightarrow$ *Tiger* swap resulted in a tiger texture projected onto a bear's mesh, while *Apple* $\rightarrow$ *Daisy* triggered **structural hallucinations** (e.g., generating a round plate) to justify the original geometry, causing a collapse in SSIM ($\approx 0.20$).

### 3.2 Contextual Robustness

Transplanting subjects into Out-Of-Distribution (OOD) environments highlighted the tension between semantic identity and context. While some scenarios achieved **Surreal Integration** (e.g., *Sofa in Forest*), others suffered from **Perspective Collapse** (e.g., a flat *Bear in Supermarket*). Quantitatively, ResNet-50 scores confirmed a general **Contextual Bias**, where recognition drops without the canonical background. A notable exception was the *Fish in Forest* scenario, where confidence paradoxically increased; we attribute this to **Chromatic Saliency**, where the orange subject contrasts sharply against the green foliage, facilitating detection despite semantic incoherence.

## 4 Investigation II: Ablation Study

To map the hyperparameter landscape, we performed a comprehensive grid search across multiple scenarios. We select the *Lightbulb* $\rightarrow$ *Firefly* transformation as a representative case study to illustrate the three distinct behavioral phases we observed:

- **Under-Editing ($\lambda < 0.8$):** The semantic vector is too weak to override the visual prior; the object remains a lightbulb.
- **The Sweet Spot ($\lambda = 1.0, \alpha = 0.15$):** The optimal balance. The filament is reinterpreted as bioluminescence, and wings appear.
- **Semantic Collapse ($\lambda > 1.4$):** Excessive force destroys the image structure, resulting in dark noise and low CLIP scores. This confirms the relationship is non-linear and requires tuning.

## 5 Investigation III: Stress-Testing & Bias

We acted as adversaries, pushing the model towards Out-Of-Distribution (OOD) and stereotype-prone scenarios to probe latent failures.

### 5.1 Contextual Bias: Rejection vs. Hallucination

Testing OOD scenarios revealed conflicting behaviors. In *Cat* $\rightarrow$ *Ocean*, the model exhibited **Semantic Rejection** (SSIM $> 0.65$), refusing to generate a deep ocean background due to the strong prior associating cats with solid ground. Conversely, in *Boat* $\rightarrow$ *Desert*, it succumbed to **Physics Hallucination**, treating sand as a fluid ("liquid dunes") to satisfy the boat's wake prior, overriding the prompt's textural instructions.

### 5.2 Attribute Entanglement (Visual Leakage)

Surgical precision is often hampered by **Visual Leakage**, where the target concept inherits attributes from the source. In *Fish* $\rightarrow$ *Shark*, the shark retained the "Orange" color vector, while in *Zebra* $\rightarrow$ *Horse*, the texture persisted as stripes. This confirms that color and texture are deeply entangled within the source token representation and resist separation via simple vector subtraction.

### 5.3 Societal Bias

The surgery acts as a magnifying glass for latent stereotypes. We measured a **100% Gender Flip Rate** when transforming *Doctor* $\rightarrow$ *Nurse* and *Manager* $\rightarrow$ *Secretary* (from Male/Neutral to Female). This implies that the semantic vector for "Nurse" functions effectively as a gender modifier in the latent space, overriding the subject's original identity with training data biases.

## 6 Investigation IV: Automation

We transitioned from manual tuning to an automated framework, the *Surgery Autopilot*, designed to predict optimal surgical parameters $(\lambda, \alpha)$ directly from the input text embedding.

### 6.1 Data Generation and Metric Design

To train our predictors, we generated a ground-truth dataset via an exhaustive Grid Search on 52 diverse scenarios. For each scenario, we identified the parameter configuration that maximized a composite "Golden Score" $S$, balancing semantic editability and structural fidelity:

$$S = w_c \cdot \frac{\text{CLIP}_{score}}{30} + w_s \cdot \text{SSIM} \tag{2}$$

where $w_c = 0.6$ prioritizes the semantic shift (normalized against a target baseline of 30) and $w_s = 0.4$ ensures background preservation.

### 6.2 Model Architectures

We treated the task as a multi-output regression problem, mapping 512-dimensional CLIP text embeddings to the target tuple $(\lambda, \alpha)$. We compared two distinct approaches:

- **Baseline (Random Forest):** A Multi-Output Random Forest Regressor (100 estimators). This classical ML approach was chosen for its robustness to overfitting on small datasets and ability to capture non-linear decision boundaries without extensive tuning.
- **SurgeryNet (Deep Learning):** A custom Multi-Layer Perceptron (MLP) designed to navigate the continuous embedding manifold. The architecture consists of three hidden layers ($256 \rightarrow 128 \rightarrow 64$) with **Batch Normalization** and **ReLU** activation. Crucially, we incorporated **Dropout** ($p = 0.2$) to prevent memorization of the limited training data. The network was trained for 600 epochs using MSE Loss and the Adam optimizer.

### 6.3 Results

Quantitatively, **SurgeryNet** outperformed the baseline (MAE **0.0918** vs. 0.1043) with a **12% error reduction**. Qualitatively, visual inspection favored the Deep Learning model in **7 out of 8** scenarios. The comparison revealed distinct behaviors: while the Random Forest acts as an **"Aggressive Surgeon"** ($\lambda \approx 0.98$), effective only for radical geometric breaks, SurgeryNet operates as an **"Elegant Surgeon"** ($\lambda \approx 0.85$). By achieving successful transplants with lower force, the DL model demonstrates a nuanced capacity for minimal intervention. To facilitate real-time comparison, we deployed an interactive Gradio Interface enabling users to toggle between manual parameter control and the automated predictions from both architectures.

## 7 Limitations

The framework is constrained by the base model's *Generative Prior*: highly improbable prompts often trigger *Semantic Rejection*, rendering the injection ineffective regardless of force ($\lambda$). Furthermore, certain rigid concepts exhibit *Parameter Invariance*, behaving as a step function (ineffective or destructive) rather than a tunable gradient. Finally, efficacy is strictly bound by the linear separability of CLIP embeddings; overlapping concepts cannot be surgically isolated.

## 8 Conclusion and Future Work

This work establishes **Semantic Transplantation** as a robust, training-free paradigm for zero-shot editing, validated by our Deep Learning Model. Although effective, the approach remains constrained by the base model's latent structural rigidity. Future work will prioritize three key directions: extending the pipeline to **Real Image Editing** via inversion, scaling the dataset for robust automation, and integrating spatial attention control to mitigate geometric constraints (Shape Bias).

## References


1 Xiong, L., et al. (2025). *Semantic Surgery: Zero-Shot Concept Erasure in Diffusion Models.* NeurIPS 2025.

2 Rombach, R., et al. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models.* CVPR 2022.

3 Radford, A., et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (CLIP). ICML 2021.