Fine-Grained Benchmarking for Image Steering

Romain Trefault (2244389) — Alban Lichet Martinez (2251896) — Clara Glotin (2249700) — Erwan Achat (2244316)

Abstract

This report details our development of a quantitative benchmark for evaluating image steering and editing models. Unlike traditional qualitative assessments that often rely on "cherry-picked" results, our goal with this framework is to provide a multi-axial evaluation of model performance using tools such as DeepFace, MediaPipe, DEtection TRansformer (DETR), Monocular Depth Estimation via a DNN (MiDaS), and Contrastive Language–Image Pretraining (CLIP) [1][2]. By measuring Physical, Structural, and Abstract changes, our system produces standardized Success Rate and Disentanglement scores, aiming to offer a transparent view into how effectively a model follows instructions without inducing unwanted side effects.

1 Introduction

The current state of image editing (e.g., FlowEdit, ControlNet) is characterized by high visual quality but low evaluative transparency. Models often suffer from entanglement, which occurs when changing one attribute (e.g., age) inadvertently changes another (e.g., background or ethnicity), but also from geometric distortion, when moving an object in 2D space results in unrealistic 3D depth, and from the well-known demographic bias, when style transfers alter skin tones or other biological features. Our goal with this project is to address these gaps by shifting the evaluation paradigm from "Does it look good?" to "Did it do exactly what was asked, and only what was asked?". In other words, we must avoid relying on subjective or non-automated human evaluators and instead focus on a systematic evaluation method that extracts the useful signal and assesses the quality of its editing [3].

2 Proposed Methodology

We developed our benchmarking suite along three distinct semantic axes, which are presented in *Table 1.* For each axis, the evaluation of image edits is handled by specific computer vision models, which allows us to be more accurate.

Axis	**Domain**	**Key Tools**	**Metrics Extracted**
Physical	Biological/Identity	DeepFace, MediaPipe	Eye size, skin tone (ITA), age, gender consistency.
Structural	Geometry/Spatial	DETR, MiDaS	Object coordinates (x, y), depth maps (z).
Abstract	Style/Intensity	CLIP Embeddings	Magnitude of style transfer, semantic alignment.

Table 1: The evaluation framework across semantic domains.

We created a structured dataset consisting of "Original vs. Edited" pairs. The data is organized by intent to test specific steering capabilities. For example, `/physical/increase_age` tests aging without altering identity, `/structure/move_left` tests 2D translation while maintaining depth/shadow consistency and `/abstract/van_gogh` tests filter strength versus content preservation.

3 Definition of the Metrics

The primary challenge in image steering is the Entanglement Problem: a model may successfully change the target attribute (Signal) but inadvertently and negatively impact unrelated features (Noise). This motivates our definition of the Steering Delta, a weighted ratio of these two forces. This Steering Delta (ΔS) quantifies the efficiency of an edit by rewarding intended movement along the active axis and penalizing movement along the Passive Axes:

$$\boxed{\Delta S = \alpha \cdot S - \beta \cdot L}$$

This formula is inspired by previous work [4]. S stands for Success Score and represents the magnitude of change in the attribute targeted by the prompt. L, for Leakage Penalty, represents the unintended variance in attributes that should remain static. α, β (typically $\alpha = 1.0, \beta = 0.5$) balance goal-achievement versus feature preservation.

3.1 Axis-Specific Signal & Noise

To compute ΔS, we decompose the image into the three distinct semantic domains presented in Section 2.

(a) For prompts targeting *Age, Expression, or Identity*, we compute the Physical Delta (ΔP). The success score is measured using DeepFace logits (specifically for age changes) or MediaPipe landmark distances (specifically for expression changes). It is defined as $S_{phys} = |P_{edited} - P_{original}|$, where P denotes the scalar physical attribute extracted from the image. Leakage is measured as skin tone drift using the Individual Typology Angle (ITA) in the CIELAB color space: $L_{skin} = \sqrt{(L^*_e - L^*_o)^2 + (b^*_e - b^*_o)^2}$, where L^* denotes perceptual lightness and b^* the blue–yellow chromatic axis; the a^* (red–green) axis is excluded to reduce sensitivity to expression-induced redness and cosmetic artifacts.

(b) For prompts targeting *Position, Scale, or Depth*, we compute the Structural Delta ($\Delta\Sigma$). The success score is computed using DETR bounding box centers $\mathbf{C} = (c_x, c_y)$ and defined as $S_{struct} = 1 - \frac{\|\mathbf{C}_{edited} - \mathbf{C}_{target}\|}{\text{max\_diagonal}}$, where $\mathbf{C}_{target}$ denotes the expected target position derived from the instruction and max_diagonal the image diagonal used for normalization. Leakage is measured as geometry breakage using depth consistency, $L_{depth} = 1 - \rho(D_{original}, D_{edited})$, where D denotes monocular depth maps estimated with MiDaS and ρ the Pearson correlation coefficient.

(c) For prompts targeting *Style, Artistic Influence, or Mood*, we compute the Abstract Delta (ΔA). The success score is measured via directional CLIP similarity and defined as $S_{clip} = \cos\Big(\vec{E}_{I_{edit}} - \vec{E}_{I_{orig}}, \vec{E}_{T_{target}} - \vec{E}_{T_{orig}}\Big)$, where $\vec{E}_I$ and $\vec{E}_T$ denote CLIP image and text embeddings, and I and T the corresponding images and prompts. Leakage is measured as perceptual degradation using $L_{percept} = \text{LPIPS}(I_{original}, I_{edited})$, where LPIPS denotes the Learned Perceptual Image Patch Similarity metric.

3.2 The Unified Scoring Algorithm

Overall, the benchmark follows a strict "Unit Test" logic to evaluate the final fidelity of an image edit. First, the prompt is classified to identify the active axis (e.g., "Make the person older" → Physical). Next, the success score (S) is computed for the active axis, while leakage (L) is measured across the remaining passive axes (Structural and Abstract). The final fidelity is then defined as:

$$\boxed{\text{Fidelity} = \frac{S}{S + \sum L}}$$

4 Implementation Details

As mentioned in the abstract and in Section 3.1, our benchmarking suite integrates state-of-the-art computer vision libraries, including `diffusers`, `transformers`, `deepface`, `mediapipe`, `lpips`, and OpenAI's `CLIP`. Our "Original" images datasets are extracted from the FFHQ dataset and the COCO 2017 dataset.

4.1 Image Steering Models

As presented in Section 2, we generated the "Edited" part of our "Original vs. Edited" image pairs dataset by using two different models, which we selected to conduct the comparative analysis presented in Section 5. The first one, InstructPix2Pix (IP2P), was used to automatically generate our main dataset. Based on `timbrooks/instruct-pix2pix` and Stable Diffusion, it employed 100 inference steps and an image guidance scale of 1.5 to achieve high-fidelity edits while maintaining structural integrity [5]. The second one, Google Gemini, was assessed using a manually annotated mini-dataset to compare its image editing capabilities against the IP2P baseline.

4.2 Evaluation Pipeline

As presented in Section 3, our evaluation method follows a pipeline that processes image pairs (Original vs. Edited) through specialized analysis models. First, the Physical axis is analyzed: facial attributes such as age and expression are detected using DeepFace, while MediaPipe extracts precise landmarks to ensure geometric consistency, and skin tone drift is calculated by converting RGB images to the CIELAB color space. Second, the Structural axis is assessed: object localization is performed using the DETR (DEtection TRansformer) model, and structural leakage is monitored via depth map correlation using the DPT (Dense Prediction Transformer) and Pearson's correlation coefficient. Third, the Abstract axis is evaluated: semantic alignment is measured with CLIP ViT-B/32 embeddings to calculate directional similarity, and perceptual degradation is quantified using the LPIPS metric with a VGG backbone to maintain visual quality.

Regarding the data processing, all images are standardized to a resolution of 512x512 pixels and converted to RGB format prior to analysis. The system uses a `benchmark_config` template to automate the processing of different semantic intents across the physical (FFHQ dataset) and structural or abstract (COCO 2017 dataset) domains.

5 Experimental Results

We conducted a comparative performance analysis between Gemini and IP2P across five standardized steering tasks. The quantitative results, detailed in *Table 2*, evaluate the models based on their ability to maximize the Steering Delta (ΔS) while minimizing leakage into passive semantic axes. Here is what we analyzed from the results of this experiment.

(a) Physical Axis (Tasks 0 & 1): In the `make_older` task, both models performed closely ($\delta \approx 0.8$). However, Gemini showed a better identity preservation. As seen in `Figure 1`, Gemini successfully introduced age-related features (wrinkles, texture) while maintaining a lower L_{skin} penalty, whereas IP2P often introduced demographic shifts.

(b) Structural Axis (Task 2): While IP2P struggled with structural collapse (Disentanglement of 0.021), Gemini maintained much higher spatial integrity (0.323), effectively translating the object without "hallucinating" background artifacts.

(c) Abstract Axis (Tasks 3 & 4): This domain showed the biggest contrast. In the `van_gogh` task, Gemini achieved a perfect Disentanglement score (1.000) and a Noise level of 0.00. *Figure 2* illustrates this effect: Gemini applies the neural style transfer as a semantic layer, whereas IP2P's edits result in "perceptual degradation," where the underlying geometry of the original image is compromised to satisfy the style prompt.

Our results indicate that Gemini consistently outperforms IP2P, in line with previous studies demonstrating that large instruction-tuned models achieve higher fidelity edits [3][5]. The choice of this comparative experiment was motivated by the desire to demonstrate the relevance of our model by testing it on datasets with known differences in editing quality. Indeed, we were unable to evaluate our method by comparing it to an existing benchmark, as it was not possible to find a model ready to serve as a reference for comparison.

On the contrary, the studies we found in the field were incorporated into our own evaluation methods. Therefore, we chose to assess our method through direct testing; however, given more time, we would have liked to confront it with a recognized reference benchmark as well.

#	Task	Axis	Winner	Gem. δ	IP2P δ	Gem. Dis.	IP2P Dis.	Gem. S/N	IP2P S/N
0	make_sad	physical	Gemini	-0.126	-0.500	0.000	0.000	0.00 / 0.25	0.00 / 1.00
1	make_older	physical	Gemini	0.816	0.800	0.731	0.714	1.00 / 0.37	1.00 / 0.40
2	move_left	structural	Gemini	-0.014	-0.029	0.323	0.021	0.29 / 0.61	0.00 / 0.06
3	pencil_sketch	abstract	Gemini	0.106	0.000	0.469	0.000	0.24 / 0.28	0.00 / 0.00
4	van_gogh	abstract	Gemini	0.140	0.136	1.000	0.785	0.14 / 0.00	0.16 / 0.04

Table 2: Comparison Results between Gemini and IP2P across Steering Delta (δ), Disentanglement metrics and Signal/Noise ratio.

Figure 1 provides a breakdown of some tasks discussed above, showcasing the Original images alongside the outputs from both competitive models. This visual comparison illustrates the generally better results of Gemini in image editing.

Figure 1: (Top) Physical Axis Analysis: Comparison of Original, IP2P, and Gemini outputs for the `make_older` task. (Bottom) Same for the Abstract Axis Analysis and the `van_gogh` style transfer.

6 Conclusion

This project established a multi-axial framework for benchmarking image steering models, establishing quantitative metrics such as the Steering Delta and Disentanglement Scores. In line with the existing documentation [1][2][3], our results demonstrate that Gemini provides better results over InstructPix2Pix, particularly in identity preservation and abstract style transfers, which constitutes a first evaluation of the relevance of our evaluation model. Future work will focus on scaling the evaluation to larger automated datasets and incorporating a reference benchmark based on existing scientific literature in order to have a reliable comparison method for better assessing the performance of our model.

7 References

1. **SpotEdit:** Ghazanfari, S., Lin, W. A., Tian, H., & Yumer, E. (2025). *SpotEdit: Evaluating Visually-Guided Image Editing Methods*. arXiv preprint arXiv:2508.18159.
2. **GIE-Bench:** Qian, Y., Lu, J., Fu, T. J., Wang, X., Chen, C., Yang, Y., Hu, W., & Gan, Z. (2025). *GIE-Bench: Towards Grounded Evaluation for Text-Guided Image Editing*. arXiv preprint arXiv:2505.11493.
3. **RISE-Bench:** Zhao, X., Zhang, P., Tang, K., et al. (2025). *Envisioning Beyond the Pixels: Benchmarking Reasoning-Informed Visual Editing*. arXiv preprint arXiv:2504.02826.
4. **PPE Framework:** Xu, Z., Lin, T., Tang, H., et al. (2021). *Predict, Prevent, and Evaluate: Disentangled Text-Driven Image Manipulation Empowered by Pre-Trained Vision-Language Model*. arXiv preprint arXiv:2111.13333.
5. **InstructPix2Pix:** Brooks, T., Holynski, A., & Efros, A. A. (2023). *InstructPix2Pix: Learning to Follow Image Editing Instructions*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).