MIND-THE-GLITCH
Visual Correspondence for Detecting Inconsistencies in Subject-Driven Generation

Abdelrahman Eldesokey, Aleksandar Cvejic, Bernard Ghanem, Peter Wonka

NEURAL INFORMATION PROCESSING SYSTEMS

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

PROJECT PAGE

PROBLEM

"<subject> leaning over a graffiti wall"

Subject-Driven Generation Model

Visual Mismatch

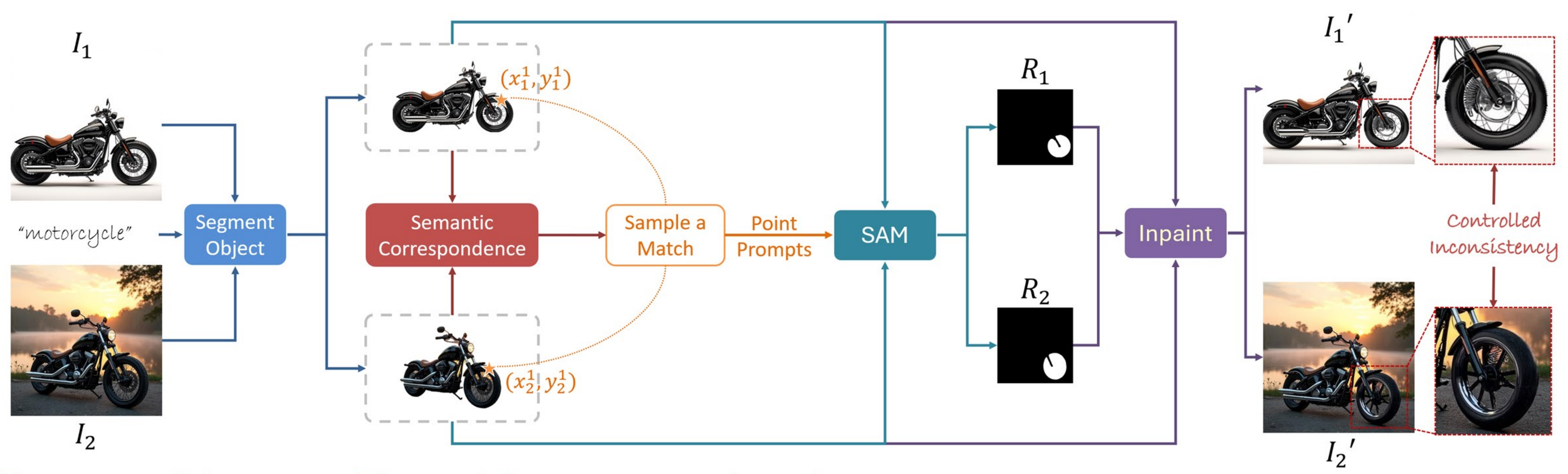
IF ONLY WE COULD MATCH IMAGES VISUALLY

We Enable Matching Images Visually

TL;DR

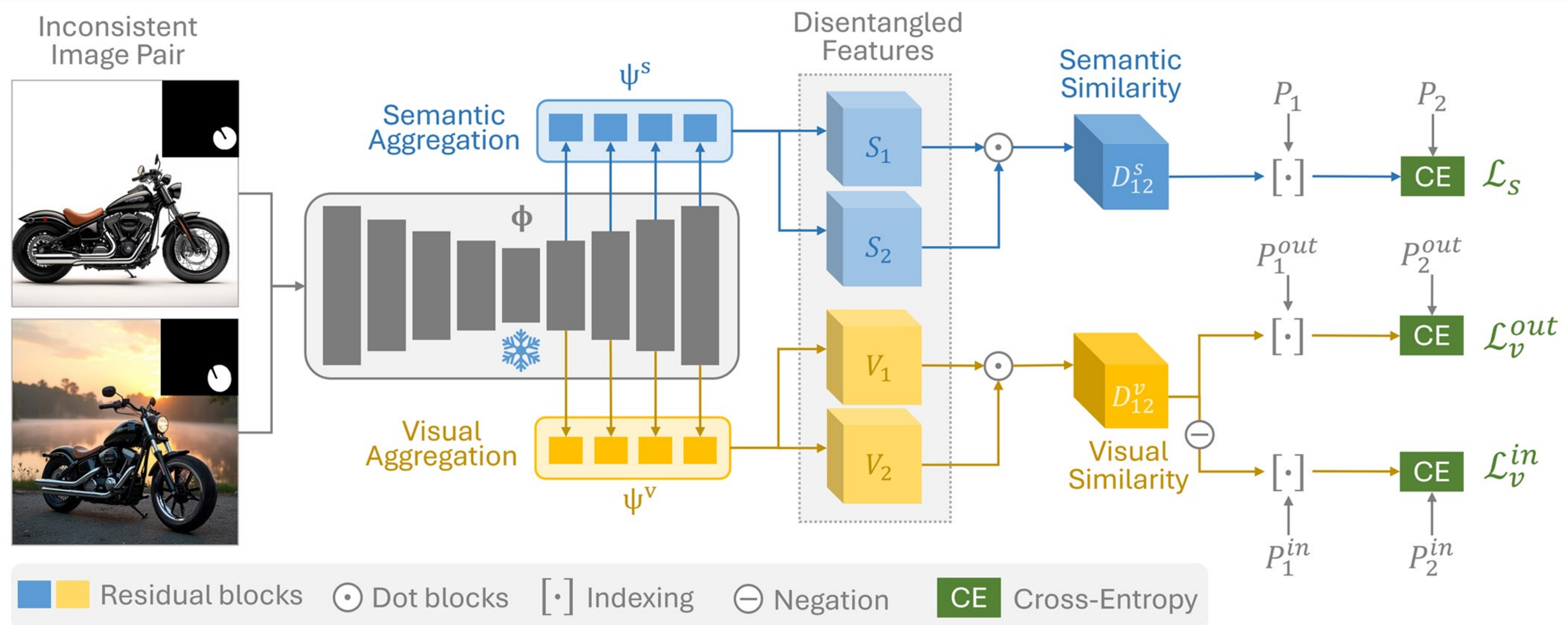
- Diffusion Model backbones *must* have visual features to support their image generation capabilities.
- We propose an approach to disentangle the features of pre-trained diffusion backbones into semantic and visual features.
- Based on these disentangled features, we derive a novel metric (VSM) that allows matching images visually.
- VSM provides a way to both quantify and localize visual inconsistencies between images supporting the evaluation of tasks such as subject-driven generation.

1. Automated Visual Inconsistency Dataset Generation



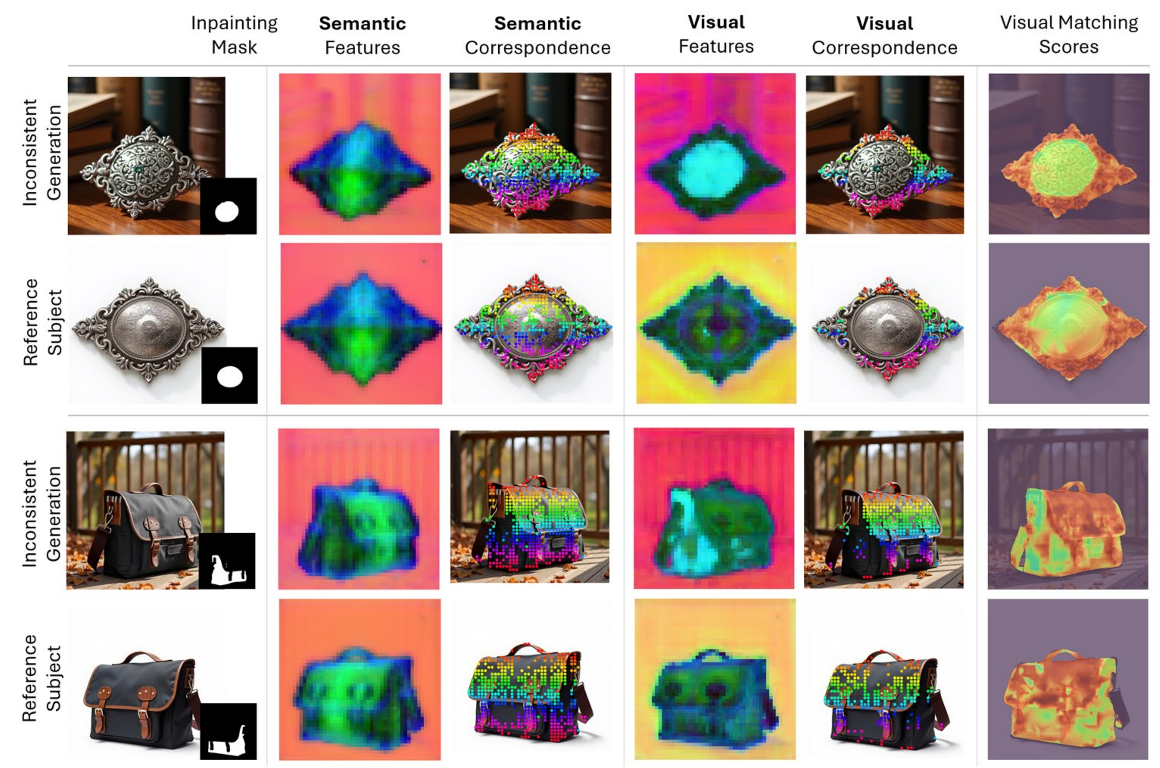
- We start with any subject-driven generation dataset.
- We visually alter (inpaint) specific parts of the subject in a controlled manner to mimic visual inconsistency.
- This produces image pairs with known visually consistent and inconsistent regions.

2. Architecture with Contastive Objective



- We use two trainable aggregation networks to extract semantic and visual features.
- Using our dataset, we pull together the features of visually similar regions and push apart the features of altered regions.
- This produces representations that are sensitive to visual changes.

3. Feature Visualization



4. The VSM Metric

$$\mathrm{VSM}(\mathcal{T}_v) = \frac{1}{|\mathcal{J}_s|} \sum_{j \in \mathcal{J}_s} \delta \left[\hat{\mathcal{D}}_j^v > \mathcal{T}_v \right]$$

The Indicator Function

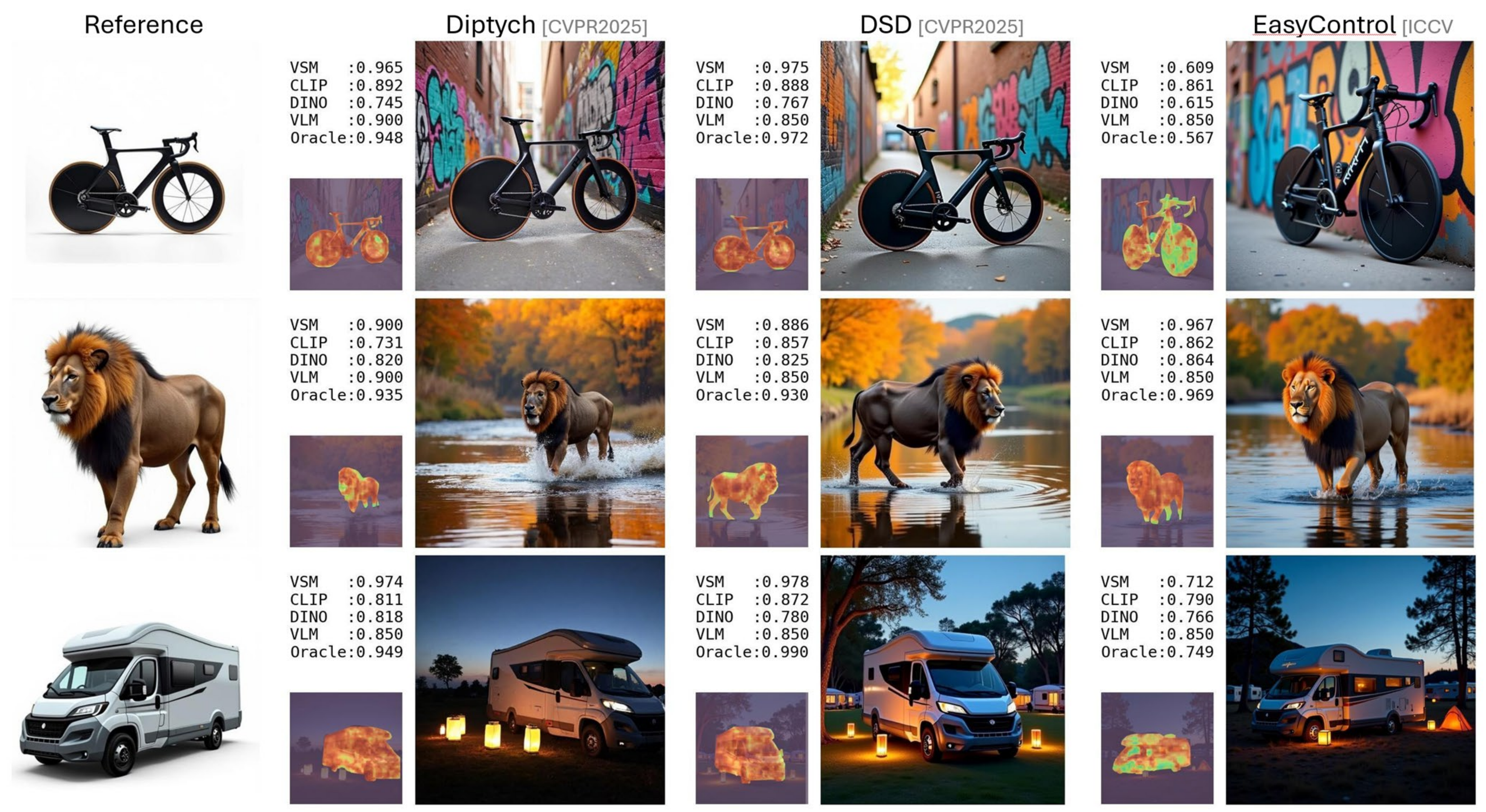
Visual Matching Score

Visual Matching Threshold

Matched Semantic Points

- We start by matching points semantically.
- For the semantically matched points, we compute the ratio of visally matched points.

5. Results on Evaluating Subject-Driven Image Generation



	Subject-Driven Generation			
	CLIP	DINO	VLM*	VSM (Ours)
Pearson	0.156	0.164	0.079	**0.405**
Spearman	0.112	0.146	0.073	**0.369**

Comparison of Score Distributions Compared to Oracle