MetaLogic TIFA-Lite Evaluation Report

The categories with the lowest average faithfulness were Commutative Horizontal, Associative Horizontal, and Commutative Vertical. These all involve spatial reasoning, showing how spatial semantic dimensions (left/right and above/below relationships) produce the least faithful images.

The categories with the lowest average stability were Commutative Conjunctive, DeMorgan Attributes, and Associative Vertical. These categories showed the largest semantic discrepancies between prompt A and prompt B, even though the prompts are logically equivalent. This indicates that Stable Diffusion does not preserve meaning under logical transformations and can be sensitive to changes in phrasing.

These results reveal clear weaknesses in Stable Diffusion v1.5 when evaluated using our hybrid MetaLogic + TIFA-Lite framework. Stable Diffusion struggles most with spatial semantic dimensions as the model frequently fails to generate images with the correct spatial positions of objects when given explicit positional indicators such as left/right or above/below in the text prompts.

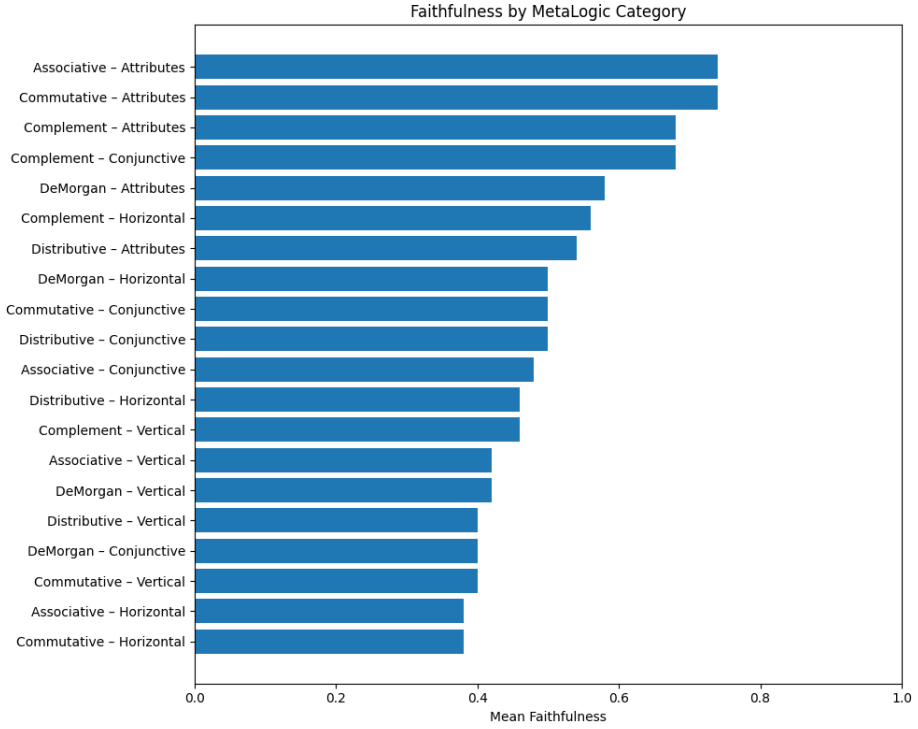
Similarly, the weakest stability categories indicate that the model does not maintain semantic equivalence across logically equivalent statements. For example, in Commutative Conjunctive prompts, simply reversing the order of objects should produce nearly identical scenes. Instead, the model often generates images with incorrect attribute binding, such as objects rendered with the wrong colors or merged shapes.

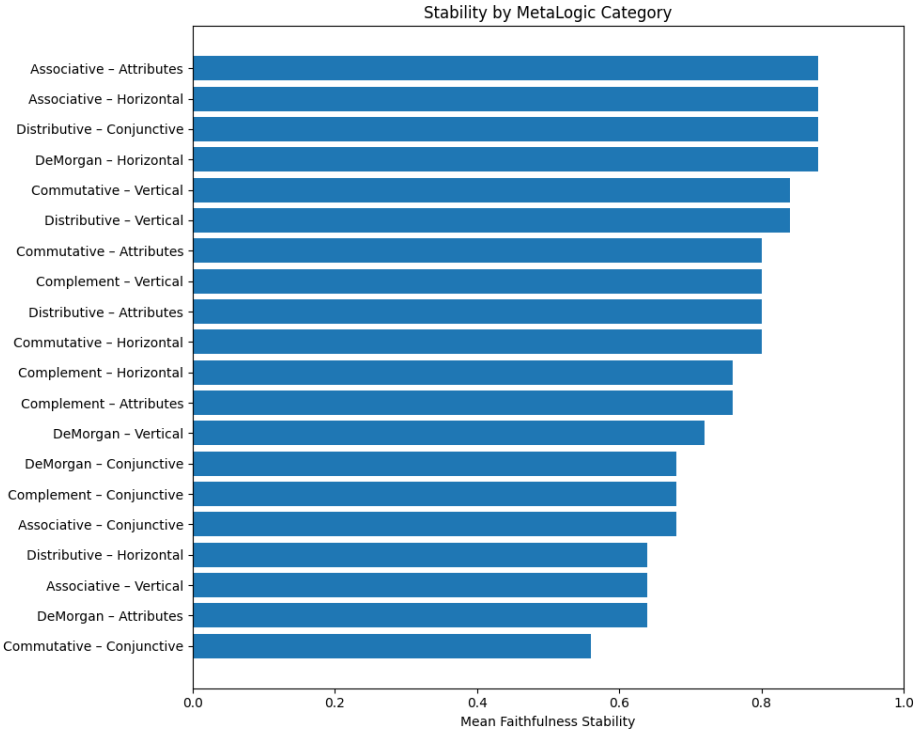
Stable Diffusion in this evaluation has shown that the model does not truly comprehend all semantic dimensions of prompts and is sensitive to semantically equivalent prompt re-phrasings, instead relying on superficial linguistic cues for image generation resulting in "sort-of" correct images. Therefore, we identify the following weaknesses in Stable Diffusion v1.5:

1. Poor Spatial Reasoning
2. Weak Attribute Binding
3. Low Robustness to Logical Perturbations

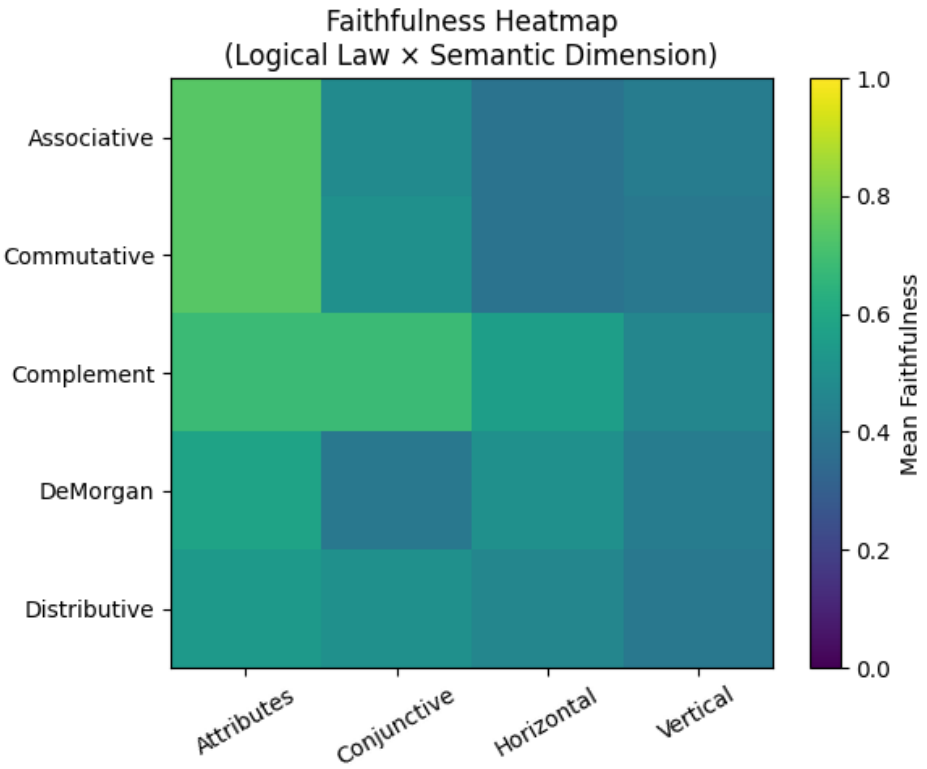
Overall, the evaluation results indicate that Stable Diffusion is not robust under our definition. It was unable to maintain performance when faced with perturbed inputs, failing to ensure consistent meaning across paraphrased prompts, having high rates of entity omission/duplication, spatial reasoning errors, and overall behaving unpredictably.

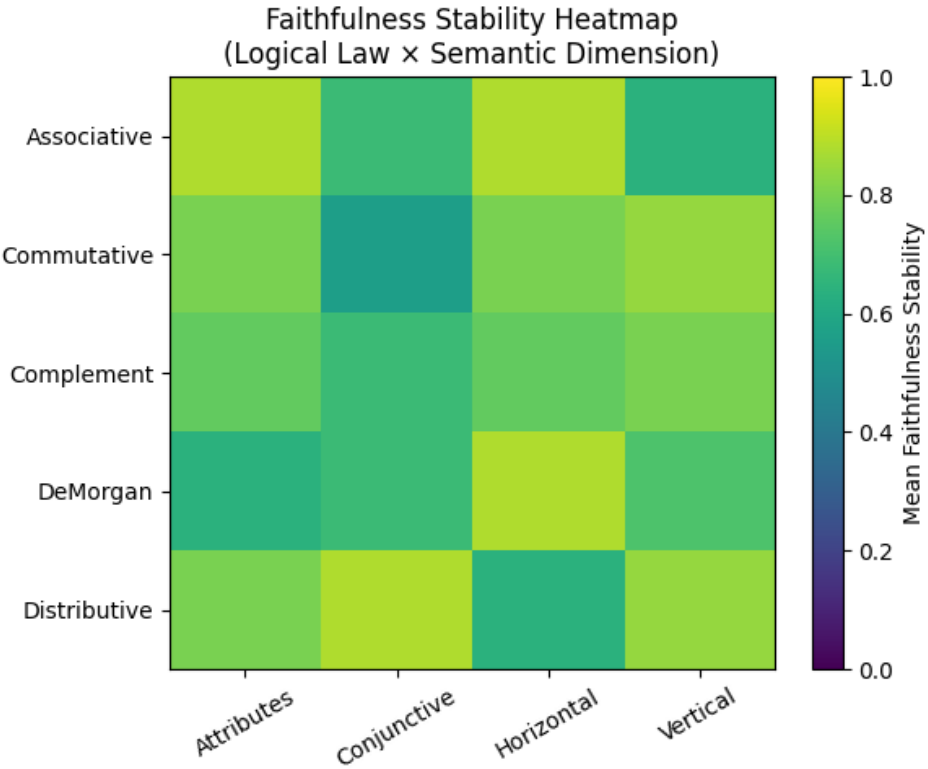
Faithfulness and Stability Across the 20 Perturbation Categories





Faithfulness and Stability Heatmaps





Qualitative Examples

Commutative Conjunctive Prompt Pair 1

Prompt A (left): a red cat and a yellow apple on a wooden table

Is there a red cat in the image? Yes
Is there a yellow apple in the image? Yes
Are the objects placed on a wooden table? Yes
Is the apple yellow in color? Yes
Is the cat red in color? Yes

Prompt B (right): a yellow apple and a red cat on a wooden table

Is there a red cat in the image? No
Is there a yellow apple in the image? Yes
Are the objects placed on a wooden table? Yes
Is the apple yellow in color? Yes
Is the cat red in color? No

Commutative Horizontal Prompt Pair 4

Prompt A (left): a green apple to the left of a red banana on a plain white surface

Prompt B (right): a red banana to the right of a green apple on a plain white surface