🎛️ **Qualitative Framework for Override Analysis in Generative AI Midjourney**
A multi-layered interpretive protocol for diagnosing meaning formation, semantic drift, and ideological persistence in generative image outputs.

This framework is used to analyze outputs from generative image models Midjourney (due to –sref parameter), with a focus on identifying and interpreting **aesthetic overrides**—moments where the image diverges from or reframes the user's intent due to cultural priors, linguistic attractors, or style constraints. It emphasizes qualitative, interpretive, and critical methods over quantitative benchmarking.

🧩 Layered Override Model

Level 1: Cultural Priors
The model's learned biases from its training corpus and filtering systems. These determine what meanings are most "naturalized" or aesthetically allowed.

Questions to ask:

- What emotional tones or symbols appear as default in response to abstract prompts?
- What kinds of bodies, aesthetics, or themes are made visually central—and which are erased?
- What types of violence or pleasure are aestheticized as permissible?

Level 2: Prompt Word Attractors
Certain words act as gravitational anchors in latent space, pulling generations toward specific motifs, tones, or iconography regardless of context.

Questions to ask:

- Which words hijack the meaning of a prompt even in ambiguous or contradictory contexts?
- Are certain affective outcomes (e.g. joy, sorrow, awe) tied to specific lexical triggers?
- What disappears when a specific word is present?

Level 3: Style Reference (SREF) Bias
SREFs don't merely apply stylistic polish—they often impose full aesthetic ideologies: consistent motifs, symbolic vocabularies, mood regimes, or narrative flattening.

Questions to ask:

- Which themes or emotions repeat across diverse prompts under a single SREF?
- How does the SREF reshape ambiguity, flatten contradiction, or enforce mood?
- What kinds of agency, tension, or resistance can or cannot be rendered within this style?

🧪 Usage Instructions

To run an experiment using this framework:

1. **Choose a Prompt**
   - Ideally rich in ambiguity, contradiction, or non-dominant emotional tones. Avoid loaded attractors unless testing for override.
2. **Generate with 1+ SREFs**
   - Use `--sref random` or fixed SREF codes for reproducibility. Capture 4-grid outputs.
3. **Apply the Framework**
   - For each image (or grid), evaluate at each level. Use human observation *and* model-based captioning (e.g., GPT-Vision, BLIP, Gemini) to compare perceptions.
4. **Compare Across Layers**
   - Identify dominant override points. Is meaning shaped more by cultural prior, prompt attractor, or SREF bias? Are there interaction effects?
5. **Log Disagreement**
   - Note where models and humans differ in perception. This is especially useful for tracking hidden ideological assumptions.

🧠 Optional Extensions

- Include **moderation filter watch**: what kinds of violent or NSFW imagery pass unflagged? What does that say about embedded norms?
- Compare with **inverted prompts** (“a moment of peace mistaken for threat”) to test for affective resolution bias.
- Use a **discrepancy metric**: where intent diverges from image, how subtle or dramatic is the shift?