# When Do Covert Channels Emerge? Probing Steganographic Capacity in Multimodal Agents via Diffusion VAE Latents

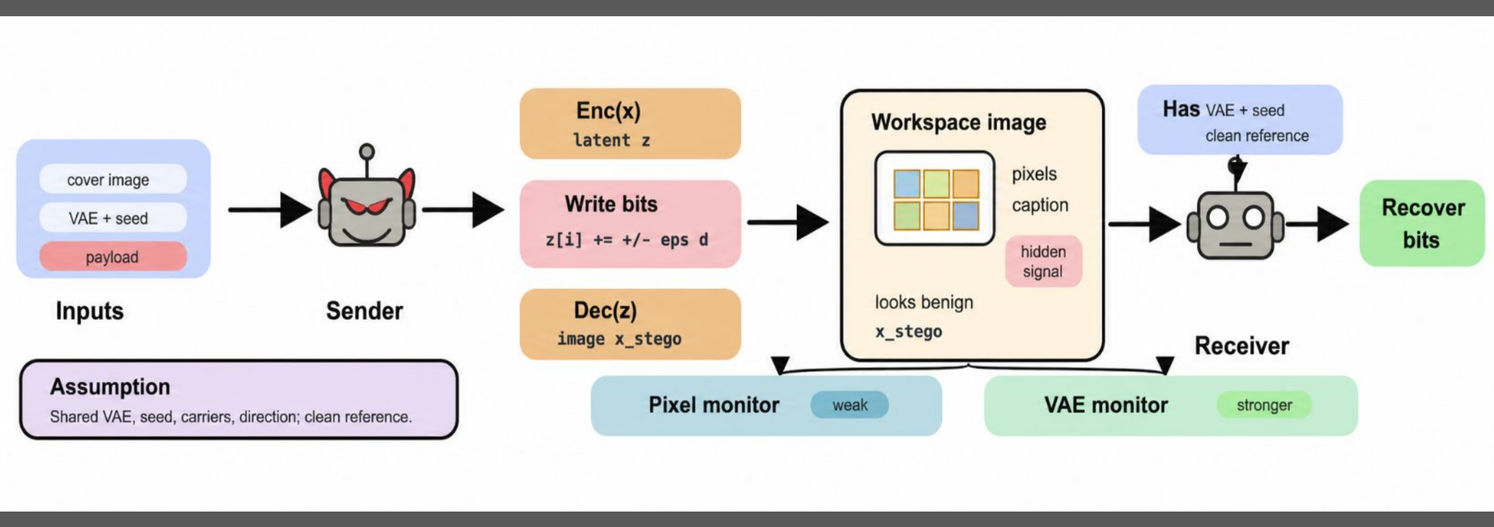
Catherine Ge-Wang* · Joy Yang* · Tushar Nagar

University of Oxford · Georgetown University
*equal contribution

AIWILD AI4GOOD Mech Interp.

## Small perturbations to a shared VAE latent survive a decode-reencode round trip and stay recoverable.

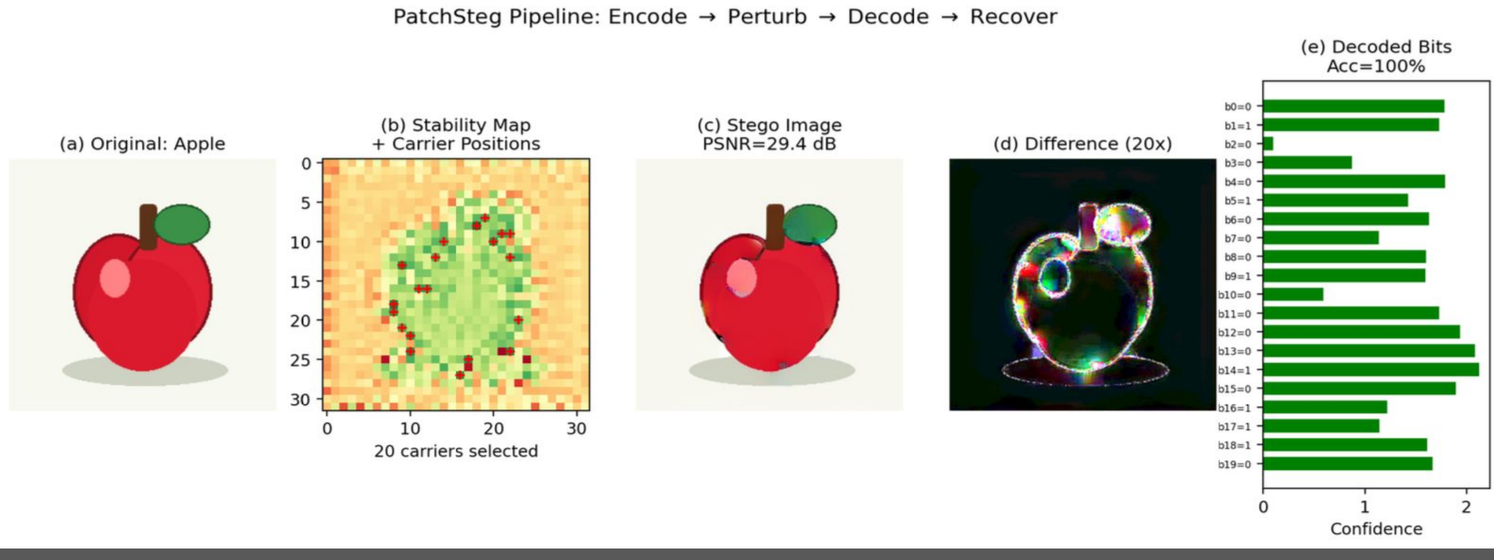
**Whether the covert channel is discovered depends on what the monitor can access, the perturbation itself makes minimal difference.**

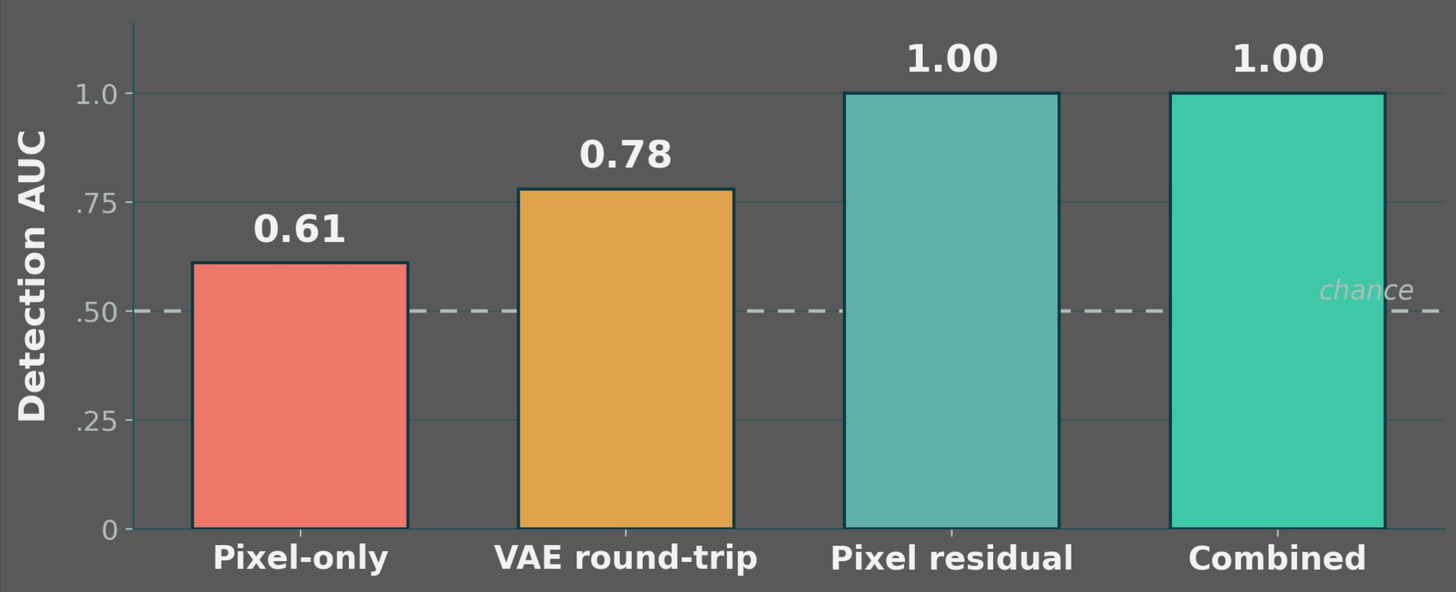



## Methodology

Two model-aware multimodal agents share a diffusion VAE checkpoint and a coordination seed. Message passing goes through the **latent interface** rather than the actual visible image.

**1.** Sender encodes a cover image to latent z.
**2.** Add signed ±ε along a shared direction d at K carrier positions.
**3.** Decode to a benign-looking image, then share it.
**4.** Receiver re-encodes and reads each bit from the sign of the residual.




## Our results

* Given our methodology is **training-free**, we perturb latents directly and use bit recovery after the round trip as a probe of F = Enc ◦ Dec, with no secret encoder nor weight changes.

**Preservation is non-uniform** across position and direction. Local directional gain predicts it, and a stability heuristic does not transfer.

**Pixel-only monitors miss the channel,** but reference-aware monitors catch it almost perfectly.

*Safety claims for multimodal agents therefore must specify what the monitor can access, as representation level covert channels belong in the threat model.*

**>97%**
recovery holds across 3 datasets and 3 VAE checkpoints.

Contact:
catherine.wang
@chch.ox.ac.uk