b. Counterfactual interventions a. Decoding the true latent z Sinusoidal Raven's PM Gene targeting Alchemy  $\overline{\phantom{a}}$  $\leftarrow$ 1.0 1.0 accuracy ( **Farget** accuracy 0.15 curacy ( W MSE ( 0.9 0.10 0.5 Relative 0.05 0.8 0.000 0.00 Graph Potion map Stone map OOD OOD OOD Latent component Implicit Explicit  $D \rightarrow$ → yq Transformer Transformer MLP Known Transformer  $x_q \rightarrow$  $\rightarrow z \rightarrow$