a. Decoding the true latent z Nonlinear reg. (MLP) Linear reg. Sinusoidal reg. Linear cls. Nonlinear cls. (MLP) 0.6 0.8 0.6 0.00100 0.004 MSE (↑ 0.6 0.00075 0.4 0.4 0.4 0.00050 0.002 0.2 0.2 ≥ 0.00025 0.2 0.00000 0.0 0.000 0.0 0.0 Raven's PM Gene targeting b. Counterfactual interventions 1.00 Rule accuracy (↑) Target accuracy ( 0.15 0.75 Alchemy Relative accuracy ( 1 0.10 1.0 0.50 0.05 0.25 0.5 0.00 0.0 Implicit Explicit  $x_q \rightarrow$ Graph Potion map Stone map Transformer  $\rightarrow y_q$ Transformer  $D \rightarrow$ MLP Transformer Known Latent component  $x_q \rightarrow$