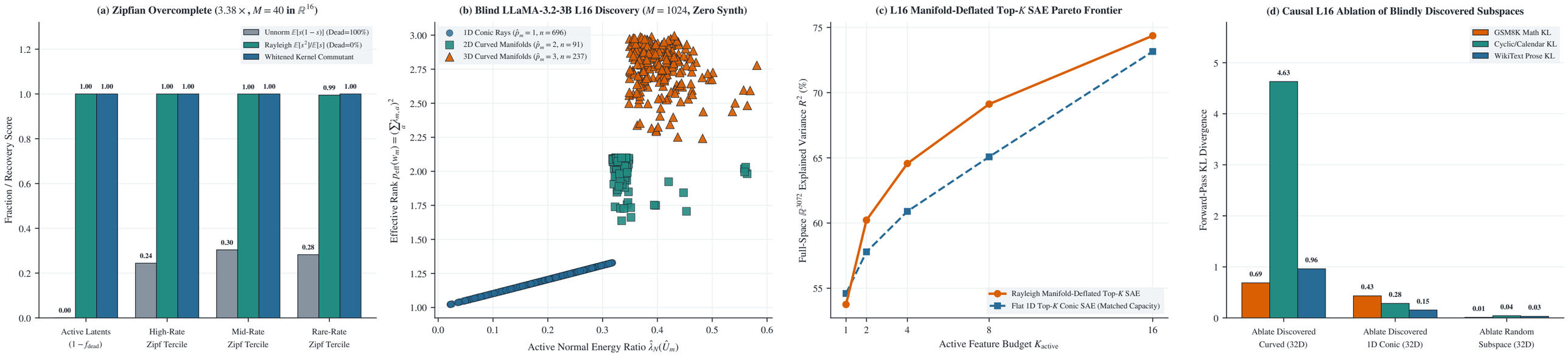

(a) Zipfian Overcomplete (3.38×, M = 40 in ℝ^16)
Unnorm 𝔼[s(1 − s)] (Dead=100%)
Rayleigh 𝔼[s²]/𝔼[s] (Dead=0%)
Whitened Kernel Commutant
Fraction / Recovery Score
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1.00
1.00
1.00
1.00
1.00
1.00
0.99
1.00
0.00
0.24
0.30
0.28
Active Latents
(1 − f_dead)
High-Rate
Zipf Tercile
Mid-Rate
Zipf Tercile
Rare-Rate
Zipf Tercile
(b) Blind LLaMA-3.2-3B L16 Discovery (M = 1024, Zero Synth)
1D Conic Rays (p̂_m = 1, n = 696)
2D Curved Manifolds (p̂_m = 2, n = 91)
3D Curved Manifolds (p̂_m = 3, n = 237)
Effective Rank p_eff(w_m) = (Σ_a λ_{m,a})² / Σ_a λ²_{m,a}
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Active Normal Energy Ratio λ̂_N(Û_m)
(c) L16 Manifold-Deflated Top-K SAE Pareto Frontier
Full-Space ℝ^3072 Explained Variance R² (%)
75
70
65
60
55
1
2
4
8
16
Active Feature Budget K_active
Rayleigh Manifold-Deflated Top-K SAE
Flat 1D Top-K Conic SAE (Matched Capacity)
(d) Causal L16 Ablation of Blindly Discovered Subspaces
GSM8K Math KL
Cyclic/Calendar KL
WikiText Prose KL
Forward-Pass KL Divergence
5
4
3
2
1
0
0.69
4.63
0.96
0.43
0.28
0.15
0.01
0.04
0.03
Ablate Discovered
Curved (32D)
Ablate Discovered
1D Conic (32D)
Ablate Random
Subspace (32D)