PPLM Vs. FUDGE Grad 0.1 love 0.1 love Linear & Softm Fixed Updated 0.6 hate 0.6 hate Latents Latents like 0.1 0.1 like \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet LLM LLM Grad Classifier P(a|x) P(xi|x1:i-1) P(xi x1:i-1) Classifier P(a|x) Multi-Head Multi-Head 0.5 love 0.1 hate like 0.4 Prompt Do you ... Prompt Do you ... \bullet \bullet \bullet \bullet \bullet