

1) Cross-Modal Retrieval

Audio

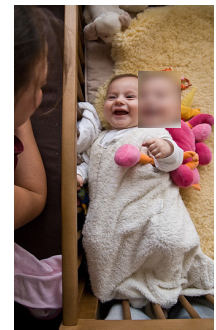


Crackle of a Fire

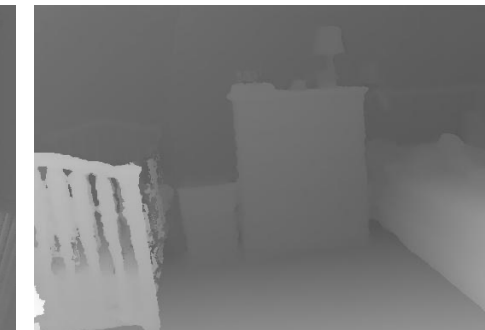
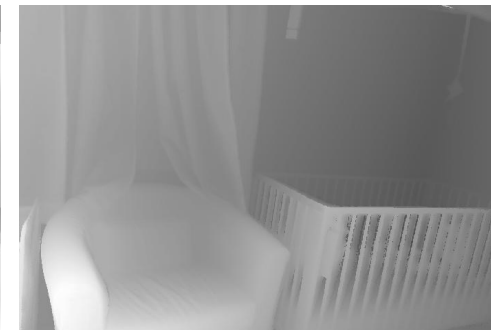
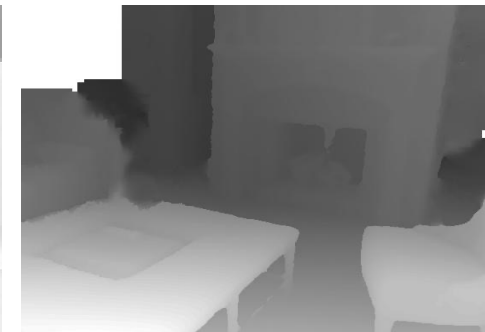
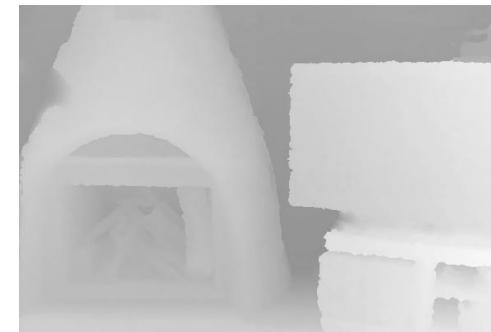


Baby Cooing

Images & Videos



Depth



Text

“A fire crackles while a pan of food is frying on the fire.”

“Fire is crackling then wind starts blowing.”

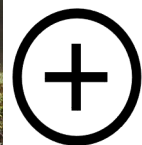
“Firewood crackles then music...”

“A baby is crying while a toddler is laughing.”

“A baby is laughing while an adult is laughing.”

“A baby laughs and something...”

2) Embedding-Space Arithmetic



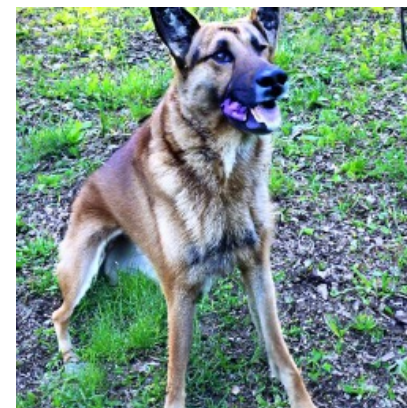
Waves



3) Audio to Image Generation



Dog



Engine



Fire



Rain

