* Ziel
  + Wir möchten in ein trainiertes Netzwerk ein neues Subject einfügen
  + Mit nur einigen Bilder wollen wir dieses Subject zu einem unique identifier binden
  + Mit Prompts wollen wir Variationen einführen und Subject in verschiedenen Kontext darstellen
* Einführung
  + Large text-to-image models lernen eine starke Prior was zum Beispiel ein Hund ist
  + Sie können aber nicht spezifische Subjects in verschiedenen Kontext darstellen (auch nicht mit super detaillierten prompts)
* Idee
  + Sie nutzen neue Prompts mit unique identifier («[V]») gefolgt von class name vom subject («dog») -> «A [V] dog»
  + Dies ermöglicht bestehende prior von class dog zu nutzen, während Instanz an unique identifier gebunden ist
* Prior work
  + Eine Möglichkeit wäre einfach finetunen
    - Man muss achten dass es nicht overfitted -> sie haben einige Papers erwähnt
* Model
  + Um «language drift» zu verhindern d.h. dass class plötzlich mit subject fotos verknüpft haben sie «autogenous class-specific prior preservation loss»
  + Prompt ist immer «A [V] dog»
    - [V] ist spezifischer Token
    - Token darf aber noch keine starke Prior im bestehenden Netzwerk haben
    - Random Text (xxy5syt) waren auch schlecht, weil tokenizer einzelne Tokens aus Buchstaben gemacht hat
    - Besser war, dass sie direkt im Vocabulary nach seltenen Tokens gesucht haben und diese dann in text space umgewandelt haben
    - -> nicht ganz verstanden
  + Class specific prior preservation loss
    - Sie finetunen alle Layers
    - Jedoch fängt model an andere Subjekte von gleicher Klasse zu vergessen
    - Dabei generiert model vorgängig selber Bilder von dieser Klasse und hat anschliessend einen Loss auf seine eigenen Bilder
    - Zusätzlicher loss (zu bestehenden noise prediction loss):
    - 
* Training
  + 3-5 Bilder mit 1000 Iterations dauerte 5 Minuten auf Nvidia A100
* Evaluation
  + Evaluate preservation of subject details in generated images
    - CLIP-I
      * Is average pairwise cosine similarity between CLIP embeddings of generated and real images
      * Ist verbreitet, kann jedoch nicht zwischen Subjekten mit ähnlichen Beschreibungen unterscheiden e.g. two different yellow clocks
    - DINO
      * Average pairwise cosine similarity between ViT-S/16 DINO embeddings of generated und real images
      * Wird bevorzugt
      * Kann zwischen Subjekten unterscheiden
  + Prompt fidelity
    - CLIP-T
      * Average cosine similarity between prompt und image CLIP embeddings
* Anwendung
  + Recontextualization
    - Subjekt in neuem Kontext
    - “a [V] teapot floating in milk”
  + Art Renditions
    - Subjekt in neuem Kunststil
    - “a painting of a [V] dog in the style of Van Gogh”
  + Novel View Synthesis
    - Ich gebe von meinem Subjekt Bilder von vorne und kann neu Bilder von hinten anhand prior erzeugen
  + Property Modification
    - Man kreuzt verschiedene Eigenschaften
    - «a cross of a [V] dog and a cat»
    - “a transparent [V] teapot”
* Ein Bild, das Text, Collage, Säugetier, Hund enthält.

  Automatisch generierte Beschreibung
* Limitations
  + Falls gewünschter Kontext sehr speziell ist failed es manchmal und kann Kontext nicht darstellen. Oder Subject ändert sich e.g. Farbe.
  + Wenn prompt sehr ähnlich von Input Images ist, kann es auch diese nehmen
  + Ein Bild, das Text, draußen enthält.

    Automatisch generierte Beschreibung