Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models

Simon Schrodi[*,1] **David T. Hoffmann**[*,1,2] Max Argus[1] Volker Fischer[2] Thomas Brox[1]





*** Equal contribution**

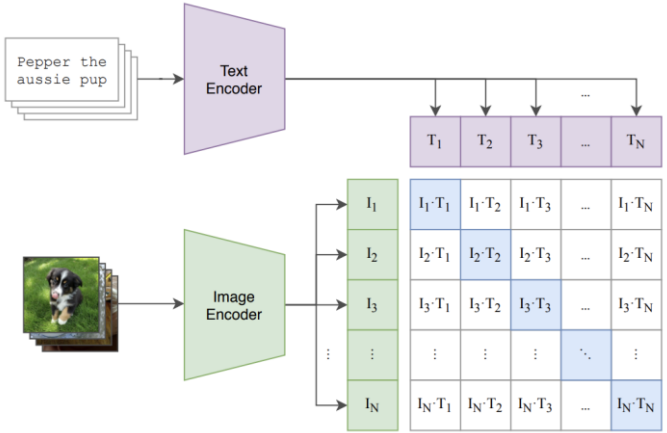
[1]University of Freiburg, [2]Bosch Center for Artificial Intelligence

universität freiburg BOSCH

DFG Deutsche Forschungsgemeinschaft German Research Foundation

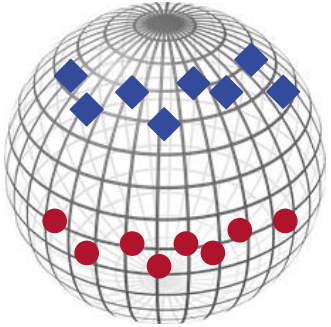
Gefördert durch: Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz aufgrund eines Beschlusses des Deutschen Bundestages

Two undesired effects of CLIP models



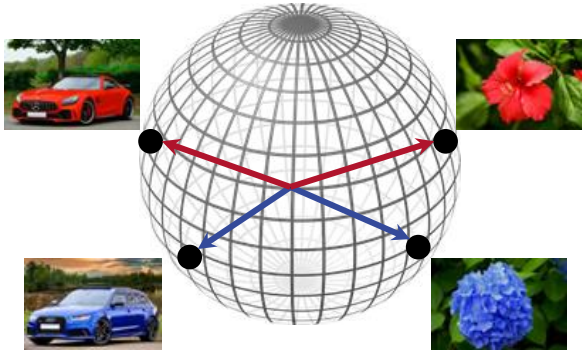
[Radford et al, ICML 2021]

Modality gap



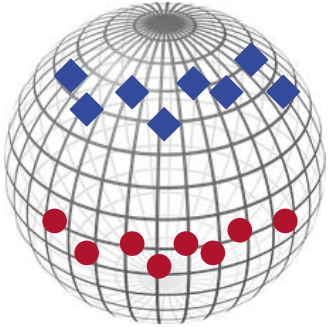
Modality A
Modality B

Object bias



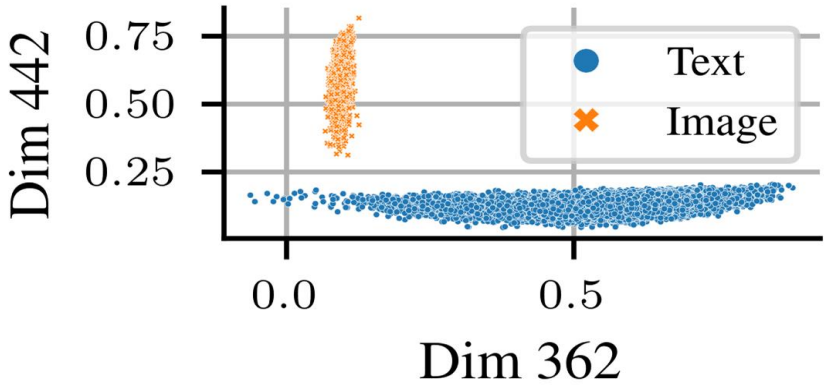
→ Common cause: Information imbalance

How does the modality gap manifest itself?

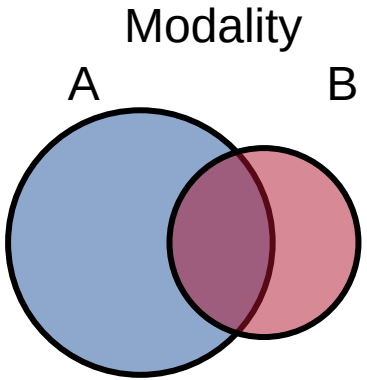


- Embeddings occupy completely separate regions of the embedding space [Liang et al, NeurIPS 2022]

- Some dimensions with high norms are primarily used by only a single modality, and vice versa:

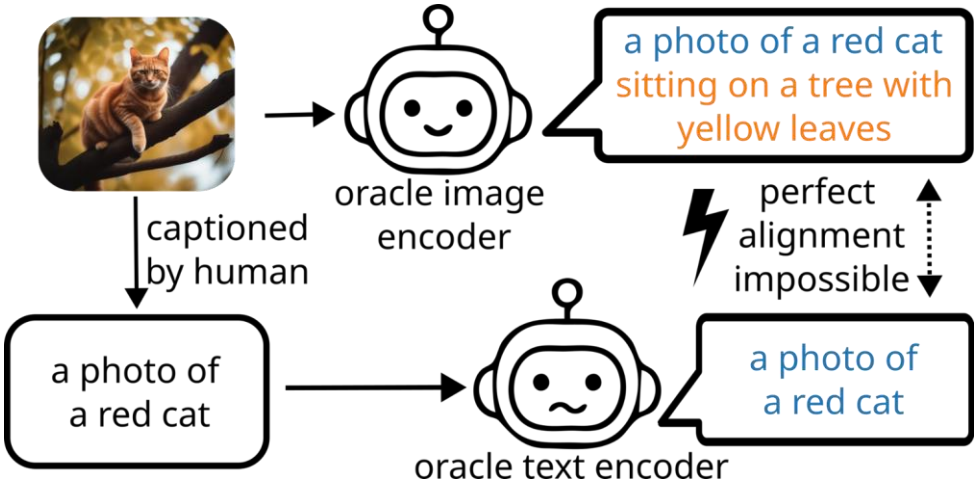


Information imbalance



- Each sample has shared information and information unique to each modality
  - It is a local phenomenon

- For example, captions typically set a focal point and ignore a lot of information



The emergence of the modality gap

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\overbrace{\log\frac{\exp(\tau f(I_i)\cdot g(T_i))}{\sum_{j=1}^{N}\exp(\tau f(I_i)\cdot g(T_j))}}^{\text{image}\rightarrow\text{text}} + \overbrace{\log\frac{\exp(\tau g(T_i)\cdot f(I_i))}{\sum_{j=1}^{N}\exp(\tau g(T_i)\cdot f(I_j))}}^{\text{text}\rightarrow\text{image}}\right)$$

- Modalities cannot be well aligned
- CLIP model is encouraged to maximize uniformity

  ⇒ For example, make the modalities as dissimilar as possible → Modality gap!

Controllable data to test our hypothesis

Real dataset

“A photo of a red cat on a tree with yellow leaves”

Synthetic MAD dataset

“2 thickening swelling no-fracture small green”

less information imbalance

Full caption: “A red cat on a tree with yellow leaves”

Half caption: “cat on a tree with yellow leaves”

Quarter caption: “cat on”

5 Attributes: “2 thickening swelling no-fracture small green”

3 Attributes: “2 swelling no-fracture small”

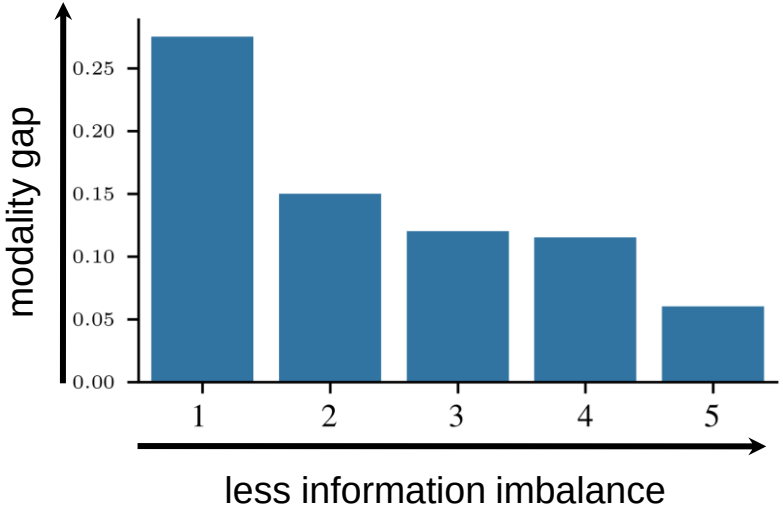
1 Attributes: “2 swelling”

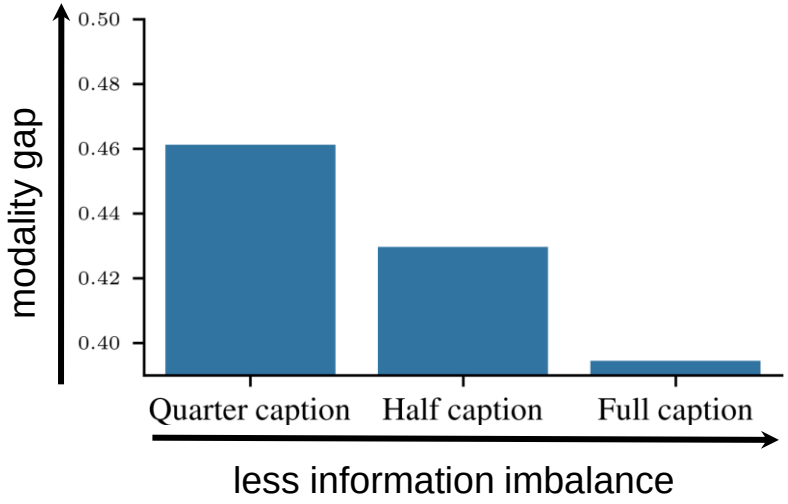
high information imbalance

Information imbalance controls modality gap

Synthetic dataset

Real dataset





But what's the purpose of the modality gap?

It might be a way of the model to adapt the entropy!

- Higher information imbalance → larger modality gap
- Higher information imbalance → higher data uncertainty
  Caption to image matching less clear
- Higher information imbalance → higher model uncertainty?

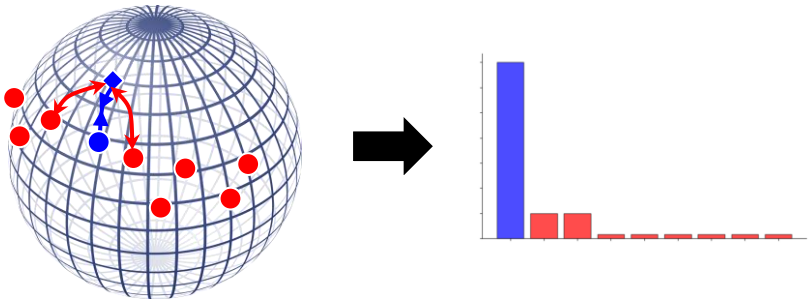
$$\mathcal{L}_{\text{CLIP}}(I,T) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\tau f(I_i)^T g(T_i))}{\sum_{j=1}^{N}\exp(\tau f(I_i)^T f(T_j))}$$

Large $\mathcal{T}$

Small $\mathcal{T}$

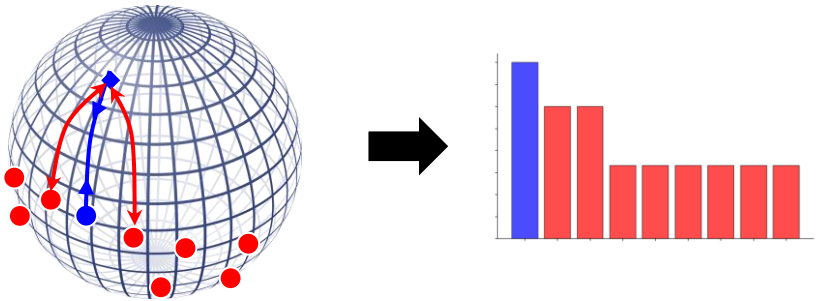
Small gap

Small entropy



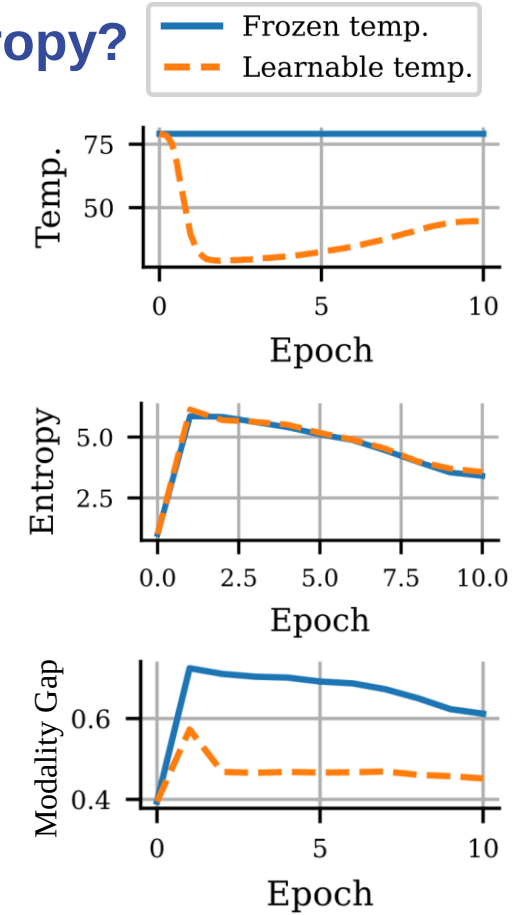
Large gap

Large entropy



Experiment: Is the modality gap linked to entropy?

1. Train model on CC12M
2. Fine-tune this model with higher information imbalance
   This data intervention increases the entropy of the data
   We finetune in two settings:
   i. With frozen temperature
   ii. With learnable temperature

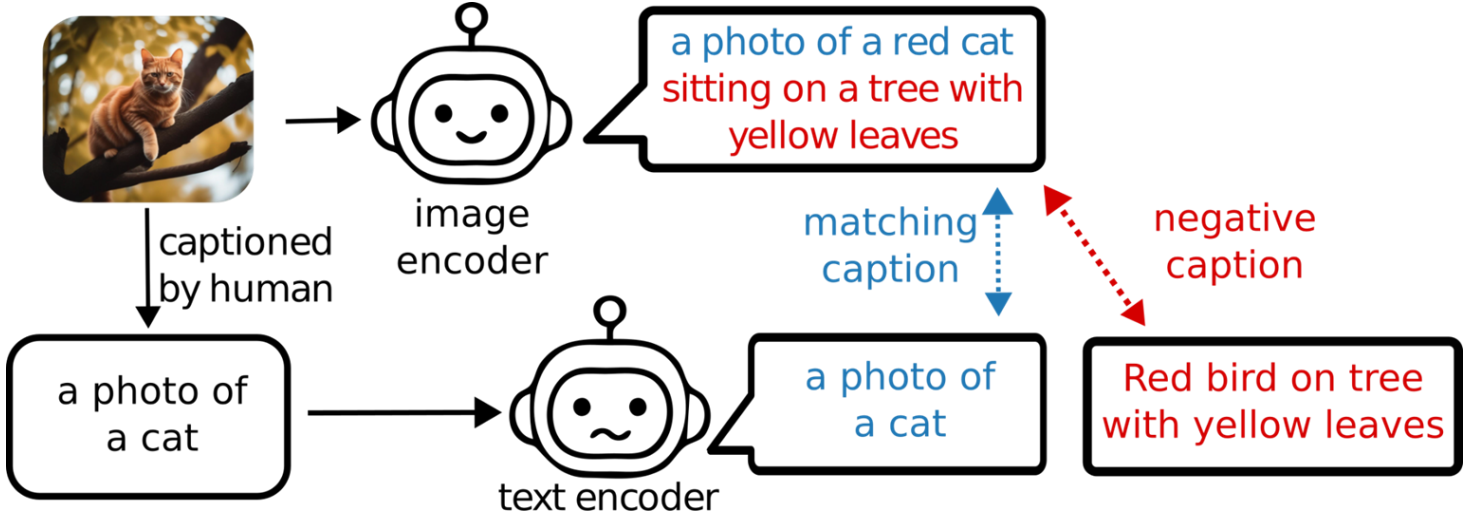


What do we expect?

1. Entropy of both models changes similarly
2. Frozen temperature: gap increases more

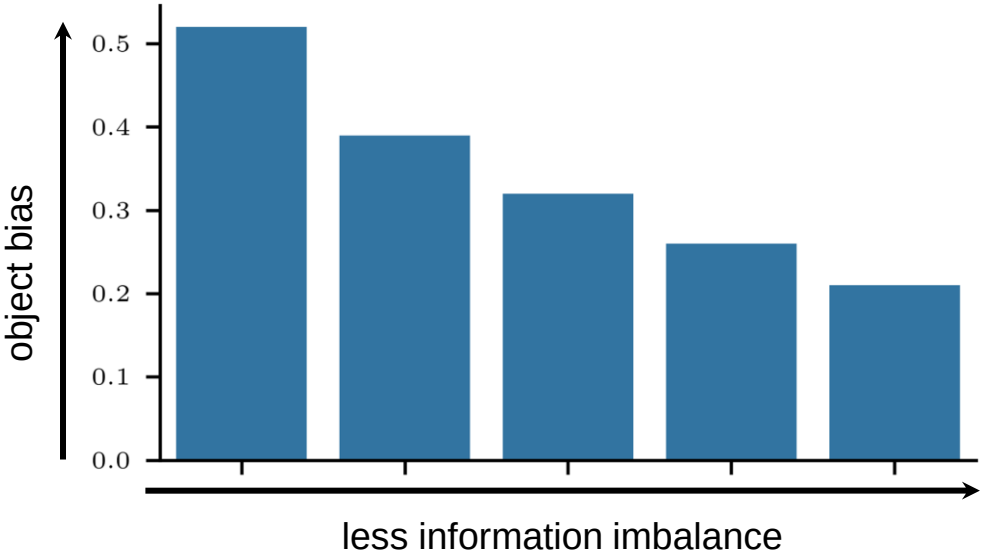
→ The modality gap could be a feature to modulate entropy of the model

Is object bias also a result of information imbalance?



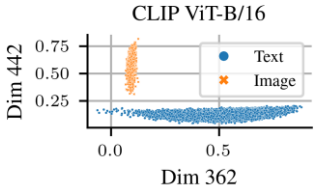
- Most humans mention central object in caption
- But mention a different set of attributes

Information imbalance controls object bias

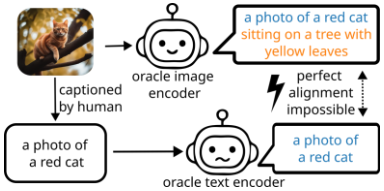


- The lower the information imbalance the smaller the object bias

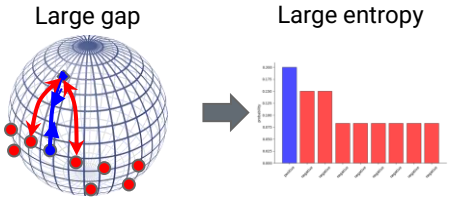
Summary



Only few embedding dimensions contribute to the modality gap



Information imbalance leads to both modality gap and object bias



Modality gap influences the entropy of the model

Thank You For Your Attention!

Paper

Code