On the Value of Cross-Modal Misalignment in Multimodal Representation Learning

Yichao Cai, Yuhang Liu, Erdun Gao, Tianjiao Jiang, Zhen Zhang, Anton van den Hengel, Javen Qinfeng Shi

November 2025

Australian Institute for Machine Learning (AIML), University of Adelaide

THE UNIVERSITY of ADELAIDE

Australian Institute for Machine Learning

Agenda

Background & Motivation

A Theory of Cross-Modal Data Misalignment

Empirical Validation

Conclusion & Broader Impact

Background & Motivation

Motivation: A Contradiction in Contrastive VLM

The Problem: Real-World Image-Text Data is Not Aligned

In real-world data, the text **t** is not a perfectly faithful representation of image **x**, which creates **cross-modal misalignment**.

Two Opposing Views on Misalignment

View 1: Harmful

- It's noise that provides weak or misleading supervision.
- It leads to model "hallucination" and poor grounding.
- **Argument:** Mitigate it, filter it.

View 2: Helpful

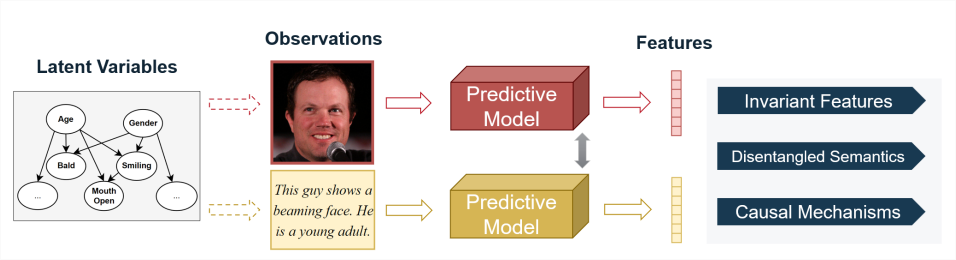
- CLIP/ALIGN are trained on noisy data yet achieve incredible robustness.
- It acts as a form of implicit regularization (e.g., our ECCV work).
- **Argument:** Understand it, leverage it.

Our Goal: A Unified Theoretical Framework

- **Core question**: How can we theoretically reconcile these two opposing views, and determine which should guide practical applications?
- **Hypothesis:** The mechanism of misalignment isn't random; it's effect depends on the context/downstream task.
- **Our approach:**
  1. Propose a Latent Variable Model (LVM) that formalizes the mechanisms of cross-modal misalignment.
  2. Theoretically analyze the objective of Multimodal Contrastive Learning (MMCL) within this LVM.
  3. Prove what is actually learned and what is discarded.
  4. Use this result to show when misalignment hurts vs. when it helps.

A Latent Generative Perspective of Representation Learning

- **Belief:** Observations (e..g, images $\mathbf{x}$, text $\mathbf{t}$) are high-dimensional realizations generated from a low-dimensional latent space.
- **Latent generative model:** $\mathbf{x} = g_x(\mathbf{s}, \mathbf{m}_x)$ and $\mathbf{t} = g_t(\mathbf{s}, \mathbf{m}_t)$, where $\mathbf{m}_{(\cdot)}$ represents non-semantic modality-specific "noise".
- **Representation learning:** Learn an encoder $f(\mathbf{x})$ that inverts this process to recover the true semantic factors $\mathbf{s}$, up to certain admissible ambiguities.



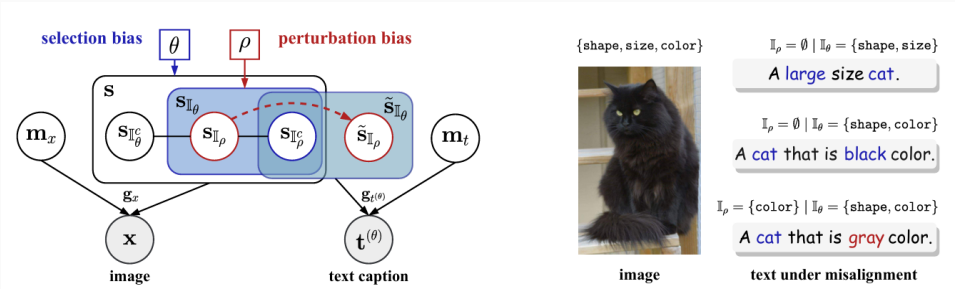
An illustration of representation learning under the generative perspective (Human face image from the CelebA-HQ dataset).

A Theory of Cross-Modal Data Misalignment

1. A Latent Variable Model for Misalignment

We formalize misalignment via two mechanisms acting on the **text generation** path:

1. **Selection bias (θ):** Characterizing Semantics are omitted from the text.
2. **Perturbation bias (ρ):** Characterizing Semantics are perturbable in the text.



An illustration of the proposed LVM for characterizing the mechanisms of cross-modal misalignment.

2. The Main Theoretical Result (Theorem 4.1)

Given this LVM, we analyze the asymptotic MMCL objective ($\mathcal{L}_{SymAlignMaxEnt}$):

$$\mathcal{L}_{SymAlignMaxEnt} = \mathbb{E}[||f_x(\mathbf{x}) - f_t(\mathbf{t})||_2^2] - \frac{1}{2}(H(f_x(\mathbf{x})) + H(f_t(\mathbf{t})))$$

Theorem 4.1: Identifiability of Latent Semantic Variables (Informal)

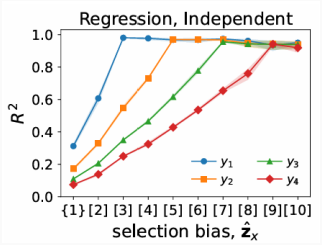
Minimizing the MMCL objective yields representations (f_x, f_t) that are **block-identifiable** with **only the subset of semantic variables that are invariant** to both selection and perturbation biases (i.e., $\mathbf{s}_{\mathbb{I}_\rho^c}$), given correct feature dimensionality.

In plain words: The model provably learns to discard any semantic factor that is omitted ($\mathbb{I}_\theta^c$) or perturbable ($\mathbb{I}_\rho$) in captions — thus, cross-modal misalignment works as a computational epistemic filter of semantic information.

3. Reconciling the Conflict & Insights (Corollary 4.1, 4.2)

When Misalignment HURTS

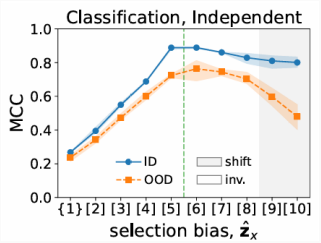
When a downstream task depends on a misaligned factor, the model discards it. Misalignment is harmful when full semantics are needed.



Misaligned task-related factors lead to poor performance.

When Misalignment HELPS

Misaligned spurious factors make the model invariant to them, acting as automatic regularization that improves OOD generalization.



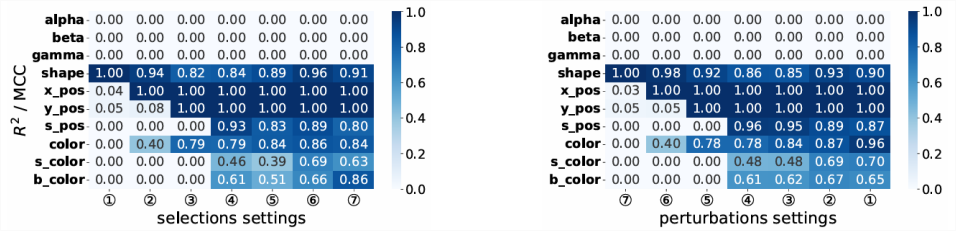
Misaligned spurious factors improve OOD generalization.

Empirical Validation

Empirical Validation 1: Synthetic Data (Causal3DIdent)

Setup: We use Causal3DIdent, which has a known latent causal graph and factors. We generate text, applying selection and perturbation biases.

Goal: Can we recover the ground-truth factors from the learned representation $\mathbf{z}_x$?



Finding: Theorem 4.1 holds

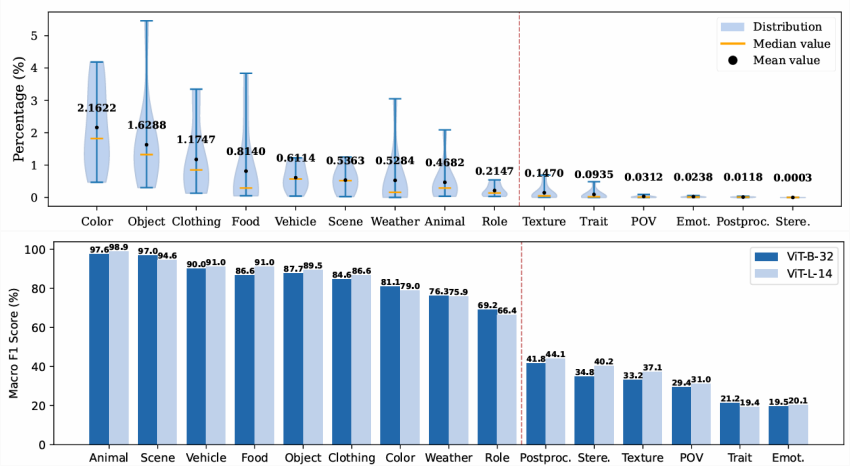
The heatmaps show a clear diagonal structure, aligning with our theoretical prediction, though extra dimensions may be predictable due to statistical correlations.

Empirical Validation 2: Real-World Case Study (OpenCLIP)

- **Question:** Does our theory hold for a real-world, large-scale model like OpenCLIP trained on LAION-400M?
- **Hypothesis (from Thm 4.1):** Concepts that are rarely captioned in LAION (i.e., high Selection Bias) should be poorly represented in the final OpenCLIP model.
- **Method:**
  1. We built a 146-concept taxonomy (e.g., "Object", "Color", "Texture", "Emotion").
  2. We measured the caption coverage (%) for each concept in LAION-400M.
  3. We measured the zero-shot Macro F1 score for each concept group using OpenCLIP.

Case Study: Findings (The "Aha!" Moment)

Observation: Concept coverage strongly correlates with zero-shot F1 performance.



Finding: Selection bias is a primary factor in multimodal representation quality.

Conclusion & Broader Impact

Conclusions & Broader Impact

1. We reconciled the two opposing views

Misalignment is not just "noise". It is a **structured signal** for representation learning.

- It **hurts** when it discards semantics you need (and when distribution shift is not an issue) (Cor 4.1).
- It **helps** when it discards spurious/sensitive semantics you **don't** need (Cor 4.2).

2. Broader Implications

- **Data Curation:** Instead of just filtering "noise", we can curate data with controlled misalignment — a principled guide for data-centric AI.
- **Ethics:** This is a epistemic mechanism for value alignment. By intentionally omitting sensitive/stereotypical concepts from captions, we can provably prevent the model from encoding them.

Thank You :)

Q & A

Project Page: https://yichaocai.com/misalignment.github.io/
Email: yichao.cai@adelaide.edu.au