Contrastive Learning: A Mathematical Perspective

Allowing Image And Text Data To Communicate

Ricardo Baptista

Statistical Sciences
University of Toronto

UNIVERSITY OF TORONTO

IMSI: Data Assimilation and Inverse Problems for Digital Twins Workshop
October 7, 2025

Collaborators

A Mathematical Perspective on Contrastive Learning

Ricardo Baptista[1], Andrew Stuart[2,3], Son Tran[3]

[1]Statistical Sciences
University of Toronto

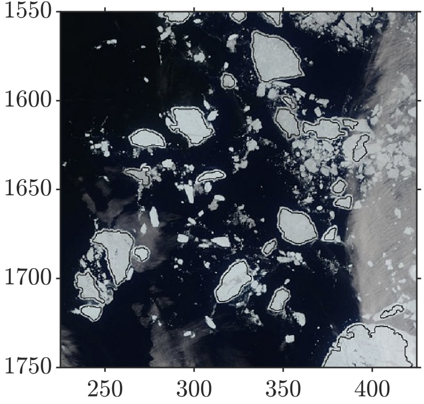
[2]Computing+Mathematical Sciences
California Institute of Technology

[3]Stores Foundational AI
Amazon

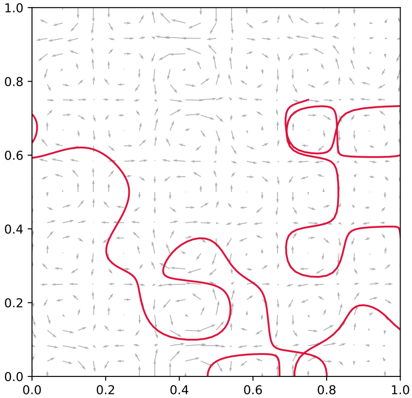
arXiv:2505.24134

Sampling Velocity Given Tracer Positions

- Infer ocean currents conditioned on data from passive tracers (e.g., sea ice floes)
- Relies on a model relating velocity to observations



Chen et al. (2021)



Sampling Images Given Text

- Generate a distribution of images conditioned on a text prompt
- For example, using Stable Diffusion XL: Podell et al. [3] (2023)
- Relies on a model to link text and images

A cat and a frog

A large aircraft taking off

An elephant in the jungle

A skateboarder in California

Aligning Text and Images

- Key contrastive learning methodology CLIP: Radford et al. [4] (2021)
- 44135 Google Scholar citations as of October 6th 2025
- Cosine similarity between $a, b \in \mathbb{S}^{\ell-1}$: $\langle a, b \rangle$
- CLIP represents text as $a \in \mathbb{S}^{\ell-1}$ and image as $b \in \mathbb{S}^{\ell-1}$: then calculate $\langle a, b \rangle$.

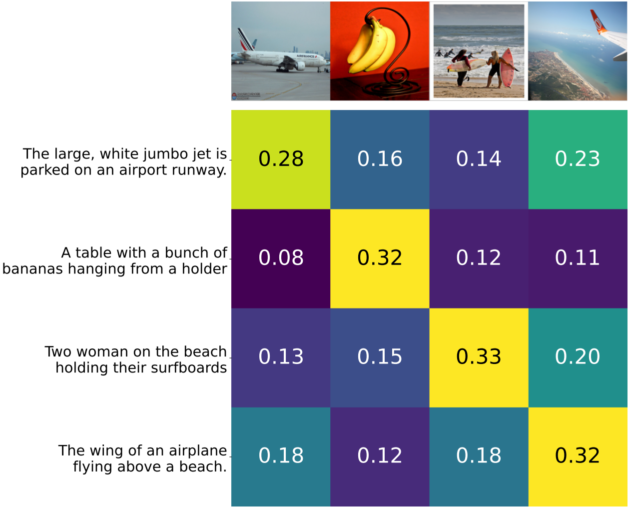


Table of Contents

Table of Contents

Setup for Contrastive Learning

Data is generated from a common reality

$$
\begin{aligned}
\textbf{Text:} \quad & u \in \mathcal{U} \\
\textbf{Images:} \quad & v \in \mathcal{V} \\
\textbf{Distribution:} \quad & \mu(du, dv) \\
\textbf{Data pairs:} \quad & \{(u^i, v^i)\}_{i=1}^N \sim \mu \text{ i.i.d.}
\end{aligned}
$$

Embed data into a common low-dimensional space

$$
g_u : \mathcal{U} \times \Theta \to \mathbb{S}^{\ell-1}
$$
$$
g_v : \mathcal{V} \times \Theta \to \mathbb{S}^{\ell-1}
$$

Encoders are represented using data-dependent architectures:

- Text: Byte-pair encoding and transformers
- Images: Convolution layers and transformers
- L^2 normalization to map to the sphere

Contrastive Learning Problem

CLIP Objective Function Radford et al [4] (2021)

Find $\theta = (\theta_u, \theta_v, \tau)$ using only samples from μ

$$
\begin{aligned}
\mathrm{L}^N_{\text{clip}}(\theta) := \frac{1}{N}\sum_{i=1}^{N}\langle g_u(u^i;\theta_u), g_v(v^i;\theta_v)\rangle/\tau \\
- \frac{1}{2N}\sum_{i=1}^{N}\log\left(\sum_{j=1}^{N}\exp(\langle g_u(u^i;\theta_u), g_v(v^j;\theta_v)\rangle/\tau)\right) \\
- \frac{1}{2N}\sum_{j=1}^{N}\log\left(\sum_{i=1}^{N}\exp(\langle g_u(u^i;\theta_u), g_v(v^j;\theta_v)\rangle/\tau)\right)
\end{aligned}
$$

$$
\theta^\star = \arg\max_{\theta}\ \mathrm{L}^N_{\text{clip}}(\theta).
$$

Objective aligned pair samples and penalizes unaligned pairs

Population-level Picture I B, Stuart and Tran (2025) [1]

Data Notation

$$
\begin{aligned}
\textbf{Data Measure:} \quad & \mu(du, dv) \\
\textbf{Data Marginals:} \quad & \mu_u(du), \mu_v(dv) \\
\textbf{Data Conditionals:} \quad & \mu_{u|v}(du|v), \mu_{v|u}(dv|u)
\end{aligned}
$$

Target Notation

$$
\begin{aligned}
\textbf{Target Measure:} \quad & \nu(du, dv; \theta) = \rho(u, v; \theta)\mu_u(du)\mu_v(dv) \\
& \quad \rho(u, v; \theta) \propto \exp\big(\langle g_u(u; \theta_u), g_v(v; \theta_v)\rangle/\tau\big) \\
\textbf{Target Conditionals:} \quad & \nu_{u|v}(du|v; \theta) = \rho(u|v; \theta)\mu_u(du) \\
& \nu_{v|u}(dv|u; \theta) = \rho(v|u; \theta)\mu_v(dv)
\end{aligned}
$$

Conditionals are weighted by the marginal data measures.

Population-level Picture II B, Stuart and Tran (2025) [1]

Population Objective Function

$$
\begin{aligned}
\mathsf{L}_{\text{cond}}(\theta) &= \frac{1}{2}\mathbb{E}_{(u,v)\sim\mu}\Big[\log \rho(u|v;\theta) + \log \rho(v|u;\theta)\Big], \\
&= \mathbb{E}_{(u,v)\sim\mu}\langle g_u(u;\theta_u), g_v(v;\theta_v)\rangle/\tau \\
&\qquad - \frac{1}{2}\mathbb{E}_{v\sim\mu_v}\log \mathbb{E}_{u'\sim\mu_u}\exp\big(\langle g_u(u';\theta_u), g_v(v;\theta_v)\rangle/\tau\big) \\
&\qquad - \frac{1}{2}\mathbb{E}_{u\sim\mu_u}\log \mathbb{E}_{v'\sim\mu_v}\exp\big(\langle g_u(u;\theta_u), g_v(v';\theta_v)\rangle/\tau\big). \\
\theta_{\text{cond}} &= \arg\max_{\theta}\ \mathsf{L}_{\text{cond}}(\theta).
\end{aligned}
$$

In practice, we minimize an empirical objective $\mathsf{L}^N_{\text{cond}} \approx \mathsf{L}_{\text{cond}}$

Theorem: Related to empirical CLIP Objective

$$
\mathsf{L}^N_{\text{clip}}(\theta) = \mathsf{L}^N_{\text{cond}}(\theta) - \log(N)
$$

For any N, the minimizers are the same, but only $\mathsf{L}^N_{\text{cond}}$ is well-defined in the population loss limit.

Population-level Picture III B, Stuart and Tran (2025) [1]

Recall: KL divergence $\mathsf{D}_{\mathsf{kl}}(\mu_1||\mu_2) = \int \log \frac{\mu_1(u)}{\mu_2(u)} \mu_1(du)$

Theorem: CLIP minimizes KL between conditionals

$$\mathsf{J}_{\mathsf{cond}}(\theta) = \frac{1}{2}\mathbb{E}_{v\sim\mu_v}\left[\mathsf{D}_{\mathsf{kl}}(\mu_{u|v}||\nu_{u|v}(\cdot;\theta))\right] + \frac{1}{2}\mathbb{E}_{u\sim\mu_u}\left[\mathsf{D}_{\mathsf{kl}}(\mu_{v|u}||\nu_{v|u}(\cdot;\theta))\right]$$

Assuming $\mu_{u|v} \ll \mu_u$ and $\mu_{v|u} \ll \mu_v$, then:

$$\arg\min_{\theta} \mathsf{J}_{\mathsf{cond}}(\theta) = \arg\min_{\theta} \mathsf{L}_{\mathsf{cond}}(\theta)$$

The non-parametric minimizer is $\nu_{u|v}(\cdot;\theta) = \mu_{u|v}$ and $\nu_{v|u}(\cdot;\theta) = \nu_{v|u}$.

Table of Contents

Generalizations of CLIP I

Alignment Metric

Joint $\nu(u, v; \theta) \propto \rho(u, v; \theta)\mu_u(du)\mu_v(dv)$ is well-defined for any ρ such that

$$\int_{\mathcal{U}\times\mathcal{V}} \rho(u, v; \theta)\mu_u(du)\mu_v(dv) < \infty$$

Examples:

- Normalized encoders: $|g_u(u; \theta_u)|_2 = 1$ and $|g_v(v; \theta_v)|_2 = 1$ with

$$\rho(u, v; \theta) = \exp\left(\langle g_u(u; \theta_u), g_v(v; \theta_v)\rangle\right) \propto \exp\left(-\frac{1}{2}|g_u(u; \theta_u) - g_v(v; \theta_v)|^2\right)$$

- Un-normalized encoders with:

$$\rho(u, v; \theta) \propto \exp\left(-\frac{1}{2}|g_u(u; \theta_u) - g_v(v; \theta_v)|^2\right).$$

Generalizations of CLIP II

Loss Function

Given a divergence D and weights $\lambda_u, \lambda_v \geq 0$:

$$J_{\mathrm{cond,D}}(\theta; \lambda_u, \lambda_v) := \frac{\lambda_u}{2}\mathbb{E}_{v\sim\mu_v}\left[\mathrm{D}(\mu_{u|v}||\nu_{u|v}(\cdot;\theta))\right] + \frac{\lambda_v}{2}\mathbb{E}_{u\sim\mu_u}\left[\mathrm{D}(\mu_{v|u}||\nu_{v|u}(\cdot;\theta))\right]$$

$$\theta^*_{\mathrm{cond}} = \arg\min_\theta J_{\mathrm{cond,D}}(\theta; \lambda_u, \lambda_v)$$

Examples:

- Choose $(\lambda_u, \lambda_v) = (1, 0)$ to match the $u|v$ conditional
- Choose divergence D computable from samples, e.g. maximum mean discrepancy:

$$\mathrm{D}(\mu_1, \mu_2) = \mathbb{E}_{(u,u')\sim\mu_1\otimes\mu_1}k(u, u') - 2\mathbb{E}_{(u,v)\sim\mu_1\otimes\mu_2}k(u, v) + \mathbb{E}_{(v,v')\sim\mu_2\otimes\mu_2}k(v, v')$$

Generalizations of CLIP III

Joint loss function

$$\mathrm{J}_{\mathrm{joint}}(\theta) = \mathrm{D}_{\mathrm{kl}}(\mu || \nu(\cdot;\theta))$$
$$\theta^*_{\mathrm{joint}} = \arg\min_\theta \mathrm{J}_{\mathrm{joint}}(\theta)$$

Implementable loss function

$$\begin{aligned}\mathrm{L}_{\mathrm{joint}}(\theta) &= \mathbb{E}_{(u,v)\sim\mu} \langle g_u(u;\theta_u), g_v(v;\theta_v)\rangle/\tau \\ &\qquad - \log \mathbb{E}_{(u,v)\sim\mu_u\otimes\mu_v} \exp\big(\langle g_u(u;\theta_u), g_v(v;\theta_v)\rangle/\tau\big) \\ \theta^*_{\mathrm{joint}} &= \arg\max_\theta \mathrm{L}_{\mathrm{joint}}(\theta)\end{aligned}$$

Objective maximizes alignment of paired data and minimizes un-alignment of un-paired data.

Empirical approximation $\mathrm{L}^N_{\mathrm{joint}}$ is more efficient to evaluate than $\mathrm{L}^N_{\mathrm{clip}}$.

Table of Contents

Low-Rank Matrix Approximations I

Goal: Analyze closed-form solutions for CLIP optimization

Setting: Gaussian data distribution $\mu = \mathcal{N}(0, \mathcal{C})$ with block covariance matrix

$$\mathcal{C} = \begin{bmatrix} \mathcal{C}_{uu} & \mathcal{C}_{uv} \\ \mathcal{C}_{vu} & \mathcal{C}_{vv} \end{bmatrix}$$

Model: Linear encoders project $u \in \mathbb{R}^{n_u}$ and $v \in \mathbb{R}^{n_v}$ into $\mathbb{R}^{\ell}$ with $\ell \leq \min(n_u, n_v)$

$$\begin{aligned} g_u(u) &= Gu, \quad G \in \mathbb{R}^{n_u \times \ell}, \\ g_v(v) &= Hv \quad H \in \mathbb{R}^{n_v \times \ell}. \end{aligned}$$

CLIP Probabilistic Model:

$$\begin{aligned} \nu(u, v; \theta) &\propto \exp\left(\langle Gu, Hv \rangle\right) \mu_u(u)\mu_v(v) \\ &= \exp\left(\langle u, Av \rangle\right) \mu_u(u)\mu_v(v), \quad A = G^\top H. \end{aligned}$$

Low-Rank Matrix Approximations II

Theorem: Minimizing Two-Sided Loss

Minimizing $\mathrm{L}_{\mathrm{cond}}$ over A with $\ell = \min(n_u, n_v)$ has solution

$$A = \mathcal{C}_{uu}^{-1}\mathcal{C}_{uv}\mathcal{C}_{vv}^{-1}$$

resulting in the conditional distributions

$$\nu_{u|v}(u|v;\theta^*) = \mathcal{N}(\mathcal{C}_{uv}\mathcal{C}_{vv}^{-1}v;\mathcal{C}_{uu})$$
$$\nu_{v|u}(v|u;\theta^*) = \mathcal{N}(\mathcal{C}_{vu}\mathcal{C}_{uu}^{-1}u;\mathcal{C}_{vv})$$

Corollary: Conditional means of $\mu_{u|v}, \mu_{v|u}$, but not variances, are recovered.

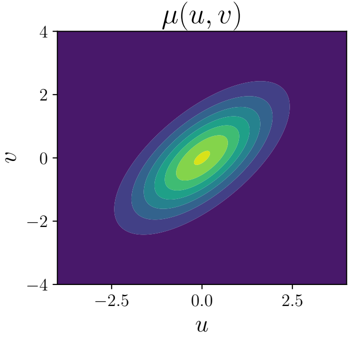
Theorem: With Dimension Reduction

Minimizing $\mathrm{L}_{\mathrm{cond}}$ over A with $\ell < \min(n_u, n_v)$ has solution

$$A = \mathcal{C}_{uu}^{-\frac{1}{2}}(\mathcal{C}_{uu}^{-\frac{1}{2}}\mathcal{C}_{uv}\mathcal{C}_{vv}^{-\frac{1}{2}})_\ell\mathcal{C}_{vv}^{-\frac{1}{2}}$$

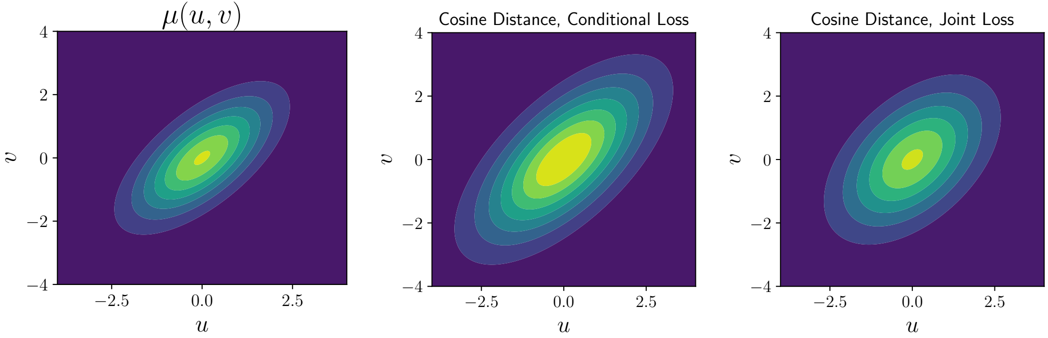
Visualizations of CLIP Generalizations I

- Two-dimensional Gaussian data distribution $\mu = \mathcal{N}(0, \mathcal{C})$
- Used un-normalized linear encoders to preserve Gaussian structure



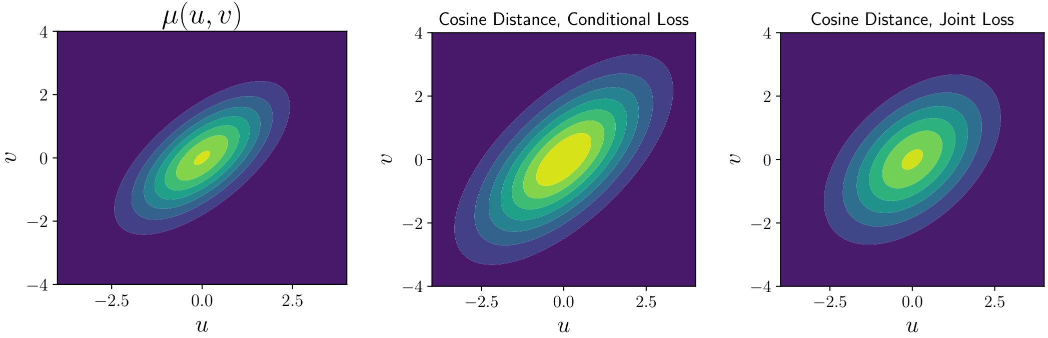
Visualizations of CLIP Generalizations I

- Two-dimensional Gaussian data distribution $\mu = \mathcal{N}(0, \mathcal{C})$
- Used un-normalized linear encoders to preserve Gaussian structure



Visualizations of CLIP Generalizations I

- Two-dimensional Gaussian data distribution $\mu = \mathcal{N}(0, \mathcal{C})$
- Used un-normalized linear encoders to preserve Gaussian structure



Takeaway:

- Conditional means, but not variance are matched with two-sided conditional loss

Visualizations of CLIP Generalizations II

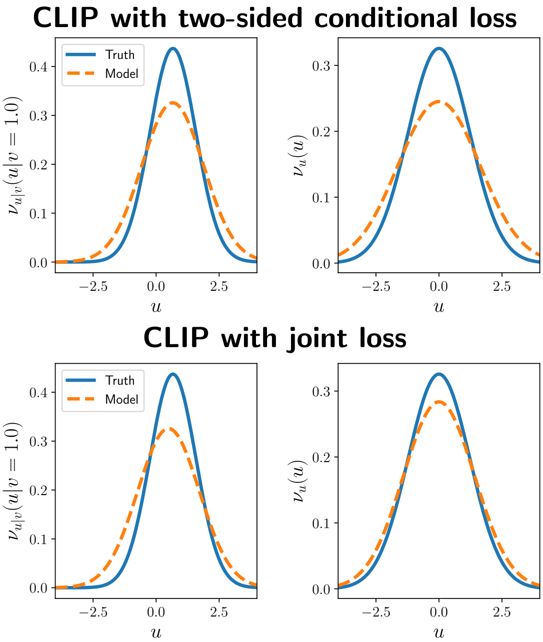
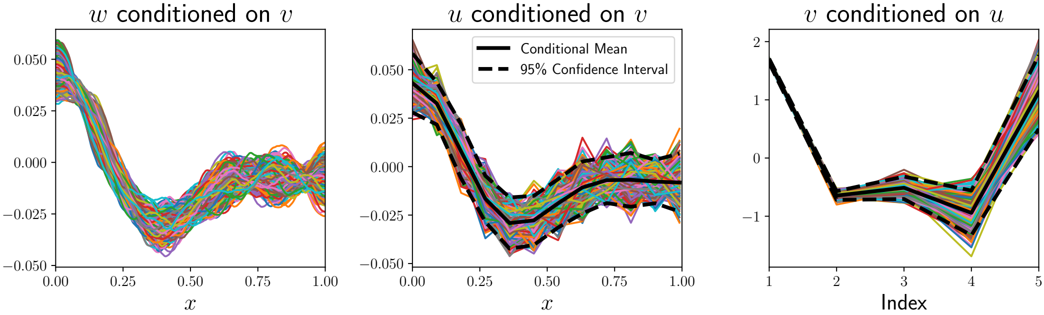


Table of Contents

Gaussian Example I

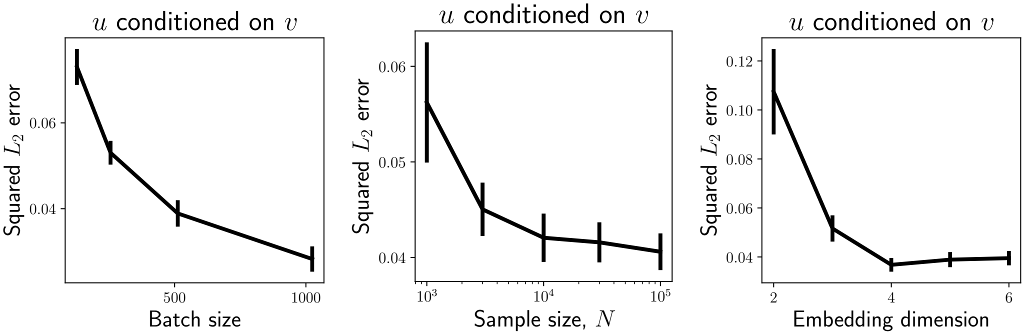
- Reality: $w \sim \mathcal{N}(0, C)$ is an un-observed Gaussian process in $L^2((0,1);\mathbb{R})$
- $u \in \mathbb{R}^{12}$ is a noisy pointwise evaluation of w at uniformly-spaced grid locations
- $v \in \mathbb{R}^5$ is five leading Fourier coefficients of w



Note: $\mu_{u|v}$ and $\mu_{v|u}$ are Gaussian with closed-form conditional means and covariances

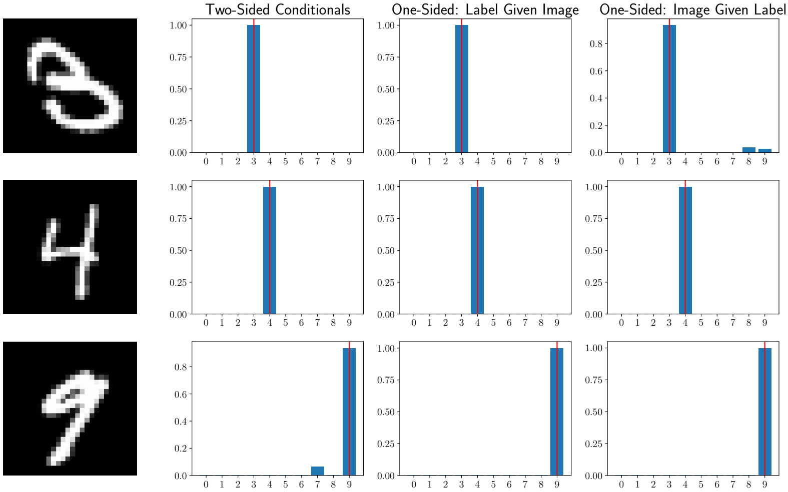
Gaussian Example II

- Used un-normalized linear encoders to represent Gaussian conditionals
- Minimized L^N_{cond} to learn encoders for both modalities
- Compared the approximation to the true conditional expectations for $u|v$ and $v|u$
- Varied the batch size, total training samples N and embedding dimension ℓ



MNIST Example

- Data modalities: images $u \in \mathcal{U} = [0,1]^{28\times 28}$ and digits $v \in \mathcal{V} = \{0,\ldots,9\}$
- Performed classifications using models learned with different loss functions

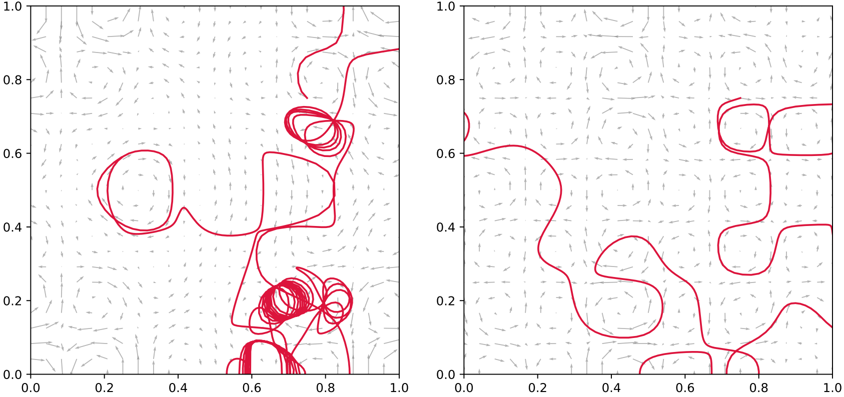


Lagrangian Data Assimilation

Original problem formulation: Kuznetsov, Ide and Jones (2003)[2]

- Velocity Field (Image); Lagrangian Trajectory (Text).
- Retrieval: identify the most likely velocity from a database given a trajectory

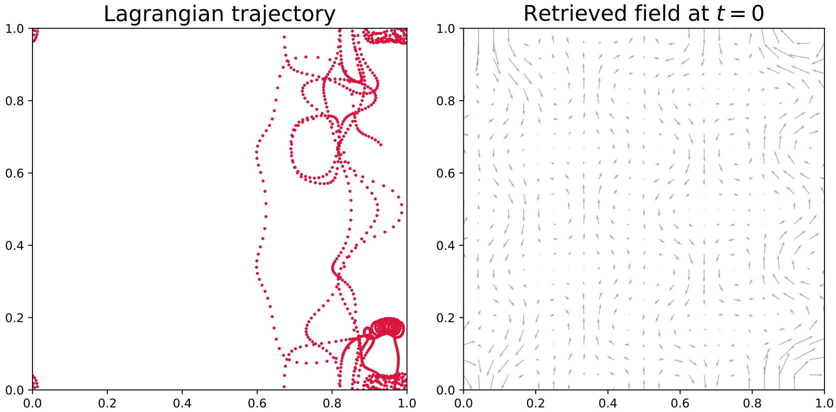
Paired potentials (background) and trajectories (red)



Lagrangian Data Assimilation

- CLIP model can align trajectory data (e.g., time-series or text) with time-dependent potential functions (e.g., images)
- Model predicts Fourier representation of the potential directly from trajectories

Paired potentials (background) and trajectories (red)



Lagrangian Data Assimilation

- Measured accuracy of retrieval between both modalities

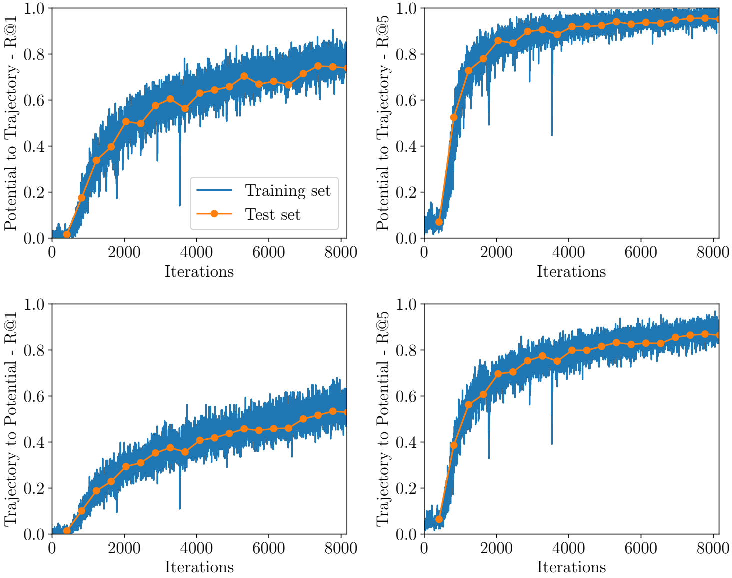


Table of Contents

Conclusions

Main Messages

- Contrastive learning relates two data modalities by tilting a product distribution
- Generalization yields new probabilistic loss functions and alignment metrics
- Encoders have closed-form low-rank matrix solutions in linear-Gaussian settings
- Application to Lagrangian data assimilation

Outlook

- Extending framework to more than two modalities
- Theoretical study of compositional generalization
- Other applications in science and engineering

References I

[1] R. Baptista, A. M. Stuart, and S. Tran.
A mathematical perspective on contrastive learning.
arXiv:2505.24134, 2025.

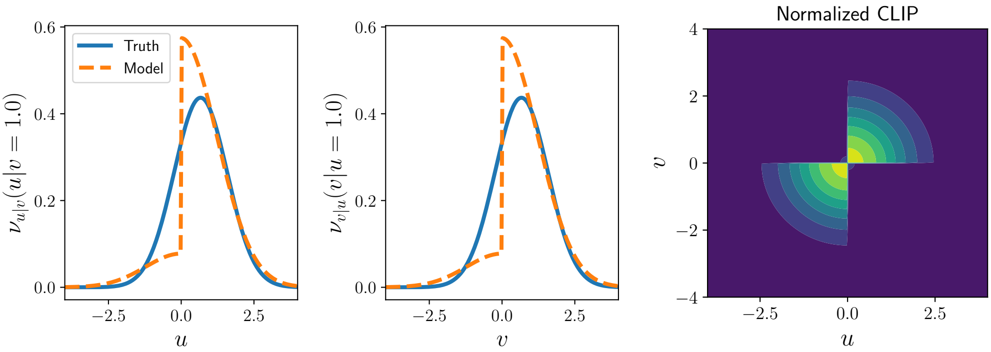
[2] L. Kuznetsov, K. Ide, and C. K. Jones.
A method for assimilation of Lagrangian data.
Monthly Weather Review, 131(10):2247–2260, 2003.

[3] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach.
SDXL: Improving latent diffusion models for high-resolution image synthesis.
arXiv:2307.01952, 2023.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al.
Learning transferable visual models from natural language supervision.
In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Visualizations of CLIP Generalizations IV

- Evaluated approximation with commonly-used normalized encoders
- Encoders capture the sign of the correlation between u and v



MNIST Example II

- Generated images conditioned on a digit using models learned with different losses
- 16 images are sampled from the conditional distribution $\nu_{u|v}$

Two-Sided Conditionals

One-Sided: Label Given Image

One-Sided: Image Given Label

Takeaway: One-sided loss for classification shows mode collapse onto a single image