ICML 2026 • Seoul

InfoAtlas:

A Foundation-style Model for *Zero-Shot* Statistical Dependency Measurement

Zhengyang Hu[†1] Yanzhi Chen[†2,3] Hanxiang Ren[4] Qunsong Zeng[1] Youyi Zheng[4] Adrian Weller[2] Kaibin Huang[1] Yanchao Yang[1,*]

[1]The University of Hong Kong [2]University of Cambridge [3]Microsoft [4]Zhejiang University [†]*equal contribution* [*]*corresp. yanchaoy@hku.hk*

300× faster than neural MI estimators

all-order linear, non-linear, non-Gaussian, discrete – all caught

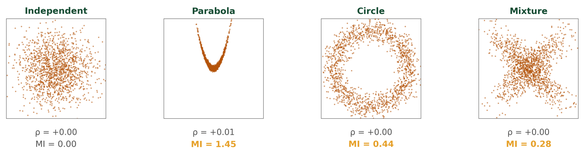
1 model handles any d_x, d_y, n – zero retraining

0 grad steps at test time – one forward pass

Why MI?

Mutual Information (MI) – the **gold-standard** measure of non-linear statistical dependence. Equivalently, the KL divergence between the joint and the product of marginals:

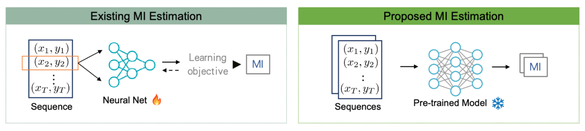
$$\mathbb{I}(\mathbf{x},\mathbf{y}) \;=\; D_{\mathrm{KL}}(p_{\mathbf{x},\mathbf{y}} \,\|\, p_{\mathbf{x}} \otimes p_{\mathbf{y}}).$$



Where MI matters. MI captures *any* dependence beyond a straight line, is invariant to per-coordinate bijections, and vanishes iff the variables are independent – the right tool whenever Pearson, Spearman, or HSIC fall short:

- ▲ feature & variable selection
- ☰ independence & CI testing
- ☒ representation analysis
- ⊛ information bottleneck
- ◇ causal discovery / ICA
- ★ image registration (medical)

Prior methods vs *InfoAtlas*



***InfoAtlas* is a *foundation-style* model pretrained on a broad family of synthetic distributions, outputting MI directly – no per-dataset retraining, no hyper-parameter tuning.**

A general-purpose, efficient toolbox

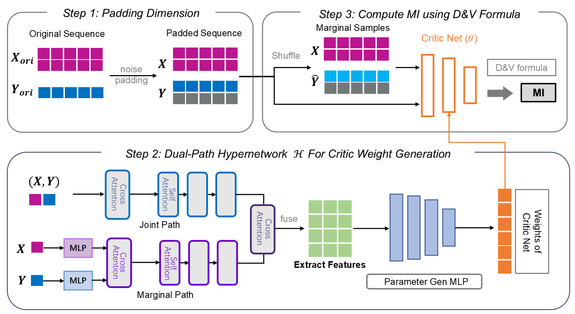
InfoAtlas ships as a **drop-in MI estimator**: feed in any dataset $\mathcal{D}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{n}$ with arbitrary d_x, d_y, n, read MI in one forward pass. **Batched inference**: 16 datasets per GPU forward, scaling near-linearly with n.

Sample size n	1 k	5 k	10 k	20 k	50 k
Time (s)	0.039	0.049	0.055	0.070	0.106
GPU mem (GB)	0.21	0.55	0.98	1.81	4.33

THEOREM

Consistency guarantee.
Under mild regularity, $\mathcal{H}$ pretrained on the synthetic meta-distribution is a **statistically consistent** estimator of $\mathbb{I}$ for any test $\mathcal{D}$ with $p(\mathcal{D}) > 0$ – asymptotic correctness without test-time gradients.

One forward pass writes the MI estimator



Key idea. A pretrained hypernetwork $\mathcal{H}$ reads an entire dataset $\mathcal{D}$ and *outputs the weights* of an MI critic in a single shot:

$$\theta = \mathcal{H}(\mathcal{D})$$

Objective. Pretrain $\mathcal{H}$ to maximise the empirical DV bound across the synthetic meta-distribution below – one D–V evaluation at test time gives MI, no optimisation.

Pretraining pipeline

Four steps per training batch, applied to fresh synthetic datasets.

Step 1 — **Sample copula mixture**: draw $(\mathbf{x},\mathbf{y})$ from a mix of Gaussian & Student-t copulas

Step 2 — **Random flow**: apply a random bijective flow to $(\mathbf{x},\mathbf{y})$ *jointly*

Step 3 — **Noise pad + softrank**: pad to common dim D; softrank → uniform margins

Step 4 — **Pretrain $\mathcal{H}$**: feed $\mathcal{D}$ to $\mathcal{H}$; train $\mathcal{H}$ to maximise empirical DV bound

Going high-dim: sliced MI

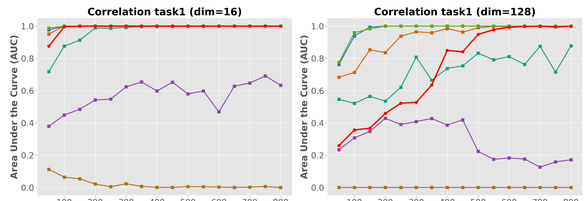
For $d \gtrsim 100$ we average MI over k-dim random projections and feed the projected pairs to the *same* $\mathcal{H}$.

$$\mathrm{SI}_k(\mathbf{x},\mathbf{y}) \;=\; \mathbb{E}_{P,P'}\big[\mathbb{I}(P^\top\mathbf{x},\; P'^\top\mathbf{y})\big], \quad k \in \{1,\dots,10\}$$

P, P' random Stiefel projectors. Sliced MI is a **theoretically grounded** estimator that preserves independence detection – $\mathbb{I}{=}0 \Leftrightarrow \mathrm{SI}_k{=}0$ – with **dimension-free** sample complexity (Goldfeld & Greenewald, 2021; Goldfeld et al., 2022).

Independence testing

Setup – one-feature linear correlation: 50 i.i.d. paired samples vs. 50 dependent samples; the estimator's MI score must rank dependent above i.i.d. (AUC).



Left: $d{=}16$ – *InfoAtlas* matches MINE / InfoNCE / MINDE without any test-time optimisation.
Right: $d{=}128$ – the zero-shot model trails optimisation-based estimators at very high d, an honest limit.

BMI benchmark (known GT MI)

Continuous benchmark of Czyz et al. (2023); $n{=}5000$ samples / task; mean over 10 seeds.

Method	Mn-dense	Spiral	Asinh@St	Student-t	Uniform	Time(s)
GT	0.59	1.02	0.45	0.18	1.02	–
KSG	0.54	0.75	0.25	0.07	0.79	0.13
MINE	**0.60**	**1.00**	0.53	0.21	**1.03**	25.9
MINDE	0.58	0.92	**0.43**	0.36	0.89	34.2
InfoNCE	0.56	0.98	0.49	**0.18**	0.97	67.6
▶ ***InfoAtlas***	**0.60**	0.89	0.41	0.21	0.93	**0.09**

300× faster than the best neural estimators on BMI, matching their accuracy.

Known-GT: non-Gaussian & discrete

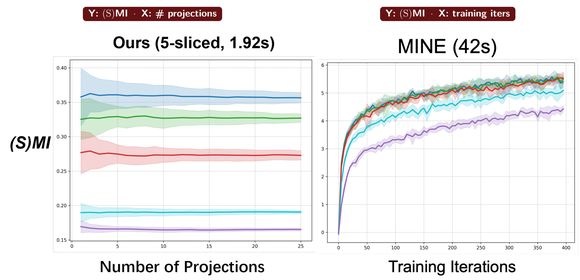
Closed-form GT MI on non-Gaussian and discrete distributions; $n{=}1000$ samples / task; lower bias is better.

Method	Beta-1d	Gamma-5d	Bern-5d	Poisson-5d	Cat-5d	Avg. Bias	Time(s)
GT	0.144	0.719	0.654	0.712	0.516	–	–
KSG	0.150	0.242	3.700	1.754	1.879	1.894	0.097
MINE	0.137	0.654	0.554	0.723	0.340	**0.130**	46.27
InfoNCE	0.135	0.502	0.606	0.725	0.233	0.201	27.00
MINDE	0.226	1.093	1.676	1.638	0.389	0.840	57.28
▶ ***InfoAtlas***	0.141	0.500	0.536	**0.573**	**0.333**	0.211	**0.046**

Generalises to **discrete and non-Gaussian** families at **>500×** the speed of MINE/InfoNCE.

CLIP image-text MI

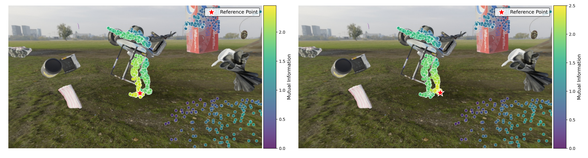
33 k COCO image-caption pairs; CLIP embeddings; **5 Gaussian-noise levels**: 0, 0.01, 0.03, 0.05, 0.1.



***InfoAtlas* (left)** separates all 5 noise levels with tight, non-overlapping bands; **MINE (right)** smears them together – at ∼100× the runtime.

3D motion segmentation

MI between **a reference point trajectory** $P^\star$ and **every other point trajectory** P on PointOdyssey; used as affinity for segmentation.



Trajectories of points on the *same* rigid object exhibit **much higher MI** ($\star$ = reference) than trajectories of points on **different** physical bodies; *InfoAtlas* recovers the segmentation in one forward pass.

Robotic key-state discovery

MI between **state** s_t and **prior state** $s_{t-\Delta t}$, *maximised* (MaxMI) to discover key states; ManiSkill2 policies trained on these key states.

Seen = training-distribution instances; Unseen = held-out test instances.

	Pick		Stack		Peg		
Method	Seen	Unseen	Seen	Unseen	Seen	Unseen	**Time(s)**
No-MI-Loss	66.6	60.0	67.4	**41.0**	38.6	9.3	–
MINE-100	86.4	81.0	68.0	37.0	55.0	13.5	0.62
MINE-1000	81.2	81.0	61.2	37.0	65.4	17.8	6.01
InfoNet	91.0	76.0	63.0	27.0	46.4	9.8	1.23
▶ ***InfoAtlas***	**94.2**	**82.0**	**68.2**	37.0	**72.4**	**18.3**	2.17

Take-home: ***InfoAtlas* — a *fast, stable, accurate toolbox* for non-linear dependence at scale. One pretrained model, any data, in real time.**

Project page

scan →