PAIRING GRAPH IDENTIFIABILITY OF MULTIMODAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

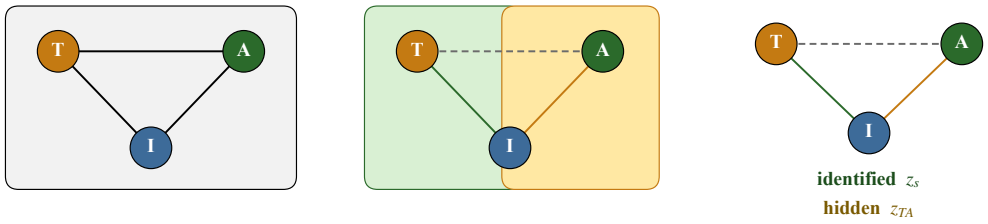
ImageBind style models do not observe a joint tuple of all modalities. They train on pairwise collections sampled independently across pairs, for instance an image and text dataset together with an image and audio dataset. Existing identifiability theorems assume the opposite observation, in which every view of a sample is present at once. They therefore do not name the latent factors that independent pairs identify. In this paper, we prove that a latent block is consistently block identifiable from pairwise datasets if and only if the modalities that carry it induce a connected subgraph of the pairing graph. Modalities are its vertices and observed pairs are its edges. On a star, this criterion identifies the factors that both spokes share with the hub, and that intersection is the content of emergent alignment. It does not identify sharing that lives only on the unobserved spoke pair. After each edge has estimated a shared representation, residual encoder disagreement equals the effective resistance of the pairing network. Among trees, a star minimises the worst pair error when the hub carries shared signal. Once that signal vanishes, spoke to spoke error returns to the prior. The identity is tight for linear Gaussian estimators, and it remains an upper bound with a nonnegative remainder for nonlinear mixing functions. Unpaired optimal transport mixes private coordinates, whereas paired barycentres do not. A content encoder of shared factor dimension leaks no private information given that factor, while an invertible embedding of a view leaks its private coordinates. Projection heads trained by edgewise InfoNCE on frozen CLIP and Qwen3-VL features recover the ranking of star, path, and disconnected graphs, and a nonlinear mixing model recovers the spoke factor only after the missing edge is added.

1 INTRODUCTION

Contrastive pretraining has made the paired sample the basic unit of multimodal learning. CLIP trains an image encoder and a text encoder with a contrastive loss on caption pairs (Radford et al., 2021), using the same family of objectives that aligns two views of one image (Chen et al., 2020). ImageBind keeps that contrastive step and changes the collection. Several image paired datasets are trained together, and text and audio then retrieve one another even though no text and audio pair was shown (Girdhar et al., 2023). LanguageBind repeats the hub pattern with language as the centre, and it binds video, audio, depth, and infrared through text rather than through images (Zhu et al., 2024). In each of these systems the learner stores pairwise datasets, and alignment appears between modalities that meet at the hub. The practice therefore raises a concrete question. Which latent factors can be recovered when the recorded joints are only those pairs, and when a sample on one pair need not be another view of the sample on a different pair.

That question already has an answer when both views of a sample are observed together, and later theory has refined the latent model while keeping this observation fixed. Two augmentations of one image are two views of a single draw. A contrastive loss isolates the content that stays fixed from the style that changes, by pulling the positive pair together and spreading the batch (Von Kügelgen et al., 2021; Zimmermann et al., 2021; Wang & Isola, 2020). The same conclusion survives the move to two modalities with distinct mixing functions. A population objective still block identifies the shared factor, and private factors stay outside a content encoder of the right dimension (Gresele et al., 2020; Daunhawer et al., 2023). Allowing many views, each mixed from its own subset of latents, yields an algebra of intersections. The likelihood is still the joint of views observed on one sample (Yao et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Simultaneously observed views. (b) Independent pairwise datasets. (c) Content identified by a star.

Figure 1: Observation models and the latent content of a star. A joint triple observes image, text, and audio together. Independent image and text and image and audio collections leave the dashed edge empty, so z_s is block identifiable and z_{TA} is not.

2024). Liu et al. (2026) recover coupled latents up to a trivial map from paired image and text, and Cai et al. (2026) show that selection or perturbation shrinks the semantic subset in that same paired regime. Each extension makes the latent description finer. The observation remains a joint law of simultaneously present views.

Hub style training removes that joint. As shown in Figure 1, a simultaneous triple is one realisation of image, text, and audio, whereas an ImageBind style collection consists of an image and text dataset together with an image and audio dataset, drawn independently. The dashed edge between text and audio is never sampled. A pair on one edge is therefore not another view of the pair on a second edge, so a selector built for $p(x_V)$ has no common tuple to read. Let $G = ([M], \mathcal{E})$ denote the undirected pairing graph, with modalities as vertices and observed pairs as edges. The algebra generated by those pairwise laws is strictly coarser than $p(x_V)$ once $M \geq 3$. A shared space learned from unpaired samples plus a few pairs still targets one joint between two modalities (Yacobi et al., 2025). The remaining gap is partial pairing. Only the edges of G are observed, and the samples used to estimate distinct edges are independent.

The identifiability theorem for this observation is a property of G. A latent block is consistently block identifiable from the edgewise contrastive losses if and only if the modalities that carry it induce a connected subgraph. On a star, emergent alignment recovers the triple intersection of the hub and both spokes, and sharing that never enters the hub is not identified. After each edge estimates a local representation, residual gauge error equals the effective resistance of the pairing network and is at most $2/\lambda_2$, where λ_2 is the algebraic connectivity (De Abreu, 2007). Among trees the star uniquely maximises λ_2, so a budget of $M - 1$ datasets should form a hub, until the hub signal to noise ratio vanishes and spoke to spoke error returns to the prior. The identity is tight for linear Gaussian estimators and is an upper bound with a nonnegative remainder for nonlinear mixers. Unpaired transport rescales private coordinates, whereas a paired barycentre does not, and flow matching inherits the split from its coupling (Gabriel & Marco, 2019; Lipman et al., 2023; Tong et al., 2024; Shrestha & Fu, 2025). A content encoder of dimension d_s leaks no private information given z_s, while an invertible embedding does, and private recovery still requires anchors (Hyvarinen et al., 2019; Locatello et al., 2020; Khemakhem et al., 2020).

We develop the observation model of independent pairwise laws, the connectivity criterion, and the resistance law for residual encoder error. They indicate which missing edge would recover a hidden block, and when a vanishing hub signal sends spoke to spoke error back to the prior. We instantiate them on linear Gaussian worlds, on a nonlinear world trained with edgewise InfoNCE, and on projection heads trained from random initialisation on frozen CLIP and Qwen3-VL features. The main contributions of this work are summarized as follows.

- A pairing graph criterion. A latent block is consistently block identifiable from pairwise datasets exactly when the induced subgraph is connected. On a star this recovers the hub intersection and leaves unobserved spoke sharing unidentified.
- A finite sample design law. Residual gauge error equals an effective resistance, so the worst pair scales as $1/\lambda_2$, tightly for linear Gaussian noise and as an upper bound otherwise. Among trees the star minimises that variance unless the hub signal to noise ratio vanishes.
- A generative and privacy counterpart. Unpaired Brenier maps mix private coordinates, whereas paired barycentres do not. Without anchors, private diffeomorphisms stay free and invertible embeddings leak private information.

2 Related work

Contrastive identifiability. Contrastive learning first isolated a content factor that stays fixed under augmentation from style that changes, since the two views are draws of one sample (Von Kügelgen et al., 2021; Zimmermann et al., 2021; Wang & Isola, 2020). The same block recovery passes to two modalities with distinct mixers, and multiview nonlinear independent component analysis is the case in which both mixers see one latent (Gresele et al., 2020; Daunhawer et al., 2023). CLIP is that two view case, so the pairing graph is a single edge (Radford et al., 2021), and composed retrieval uses the same chart for a manipulated query (Tang et al., 2026). Many views, each mixing a subset of latents, then yield an algebra of intersections, yet every selector still reads views from the same sample (Yao et al., 2024). Later work changes the latent model. Liu et al. (2026) recover coupled variables up to a trivial map from paired image and text, and Cai et al. (2026) show that selection or perturbation shrinks the semantic subset of that objective. The observation stays a simultaneous joint. A pairing graph replaces that joint by pairwise laws on independent samples, so a selector that needs $p(x_V)$ cannot be run.

Unpaired maps and missing views. A second line relaxes how views are recorded rather than which latents they mix. Deep canonical correlation analysis fills missing entries in a multiview table, so an incomplete row remains one damaged joint sample (Andrew et al., 2013). A shared space built from unpaired samples and a few pairs still targets one joint between two domains (Yacobi et al., 2025). Generative training then changes the recovered object. Brenier's theorem supplies a unique quadratic map when the source is absolutely continuous, and the map aligns the marginals (Gabriel & Marco, 2019). Flow matching learns the velocity of an interpolant, so the population field follows the coupling that builds the path (Lipman et al., 2023; Tong et al., 2024). An independent coupling can rescale private coordinates, and only a paired coupling tags the shared factor. A translation identified on unpaired domains may therefore exist and still mix those coordinates (Shrestha & Fu, 2025). When the observed pairs are edges of a graph, leftover encoder error tracks the algebraic connectivity of that graph.

Spectral glue and private factors. After each edge estimates a relative chart, a global chart is a synchronization problem on the pairing graph, as in consensus on a shared state (Qiu et al., 2026a;b). Angular synchronization and the connection Laplacian control that agreement for rotations (Singer, 2011; Bandeira et al., 2013). The rate depends on the algebraic connectivity of the graph (De Abreu, 2007). Private coordinates lie outside this glue. Nonlinear independent component analysis is impossible without an auxiliary variable (Hyvarinen et al., 2019; Locatello et al., 2019). Anchors and sparse perturbations supply that variable for factors the contrastive loss does not see (Locatello et al., 2020; Khemakhem et al., 2020; Lachapelle et al., 2022), within the broader programme of causal representation learning (Schölkopf et al., 2021). Hub training remains outside these guarantees. ImageBind aligns text and audio from image paired collections that omit the text and audio pair, and LanguageBind moves the hub to language (Girdhar et al., 2023; Zhu et al., 2024), as do federated graph backbones (Zhu et al., 2026). Neither result names the latent block the hub recovers. This paper treats the hub as a pairing design and characterises that block, the resistance of the leftover error, and the failure. The same characterisation covers any graph of pairwise datasets and multitask fusion across views (Lou et al., 2026).

3 Setup

Let $z_s \in \mathcal{Z}_s \subseteq \mathbb{R}^{d_s}$ be a factor shared by every modality. For each nonempty $A \subseteq [M]$ with $|A| \geq 2$, a block $z_A \in \mathcal{Z}_A \subseteq \mathbb{R}^{d_A}$ enters only the modalities in A. Modality m also carries a private factor $z_{p,m} \in \mathcal{Z}_{p,m} \subseteq \mathbb{R}^{d_m}$. Write $z^{(m)}$ for the concatenation of every block that enters m. The observation on that modality is

$$X_m \;=\; f_m\big(z^{(m)}\big) \in \mathcal{X}_m \subseteq \mathbb{R}^{D_m}. \tag{1}$$

Equation equation 1 is the partial observability model of Yao et al. (2024). Each view depends only on its own latent blocks. In that work the views of one sample are recorded together, so the learner sees $p(x_V)$. That joint is never observed here.

Assumption 1 (Regular mixing)**.** *Each f_m is a diffeomorphism from an open simply connected product onto its image. The latent density p_z is smooth and strictly positive on a nonempty open set.*

The dimensions of the content blocks are known. Conditional on the shared coordinates, each private factor is fully supported. This last condition is the sufficient variability assumption of Daunhawer et al. (2023).

Which of these joints are collected is a design choice, recorded by an undirected graph $G = ([M], \mathcal{E})$. For each edge $\{i, j\} \in \mathcal{E}$ the learner observes n_{ij} independent draws from the pairwise law P_{X_i,X_j}, and draws on distinct edges are independent. No sample is a triple (X_i, X_j, X_k). Let $\mathcal{F}_G$ denote the σ algebra generated by the laws $\{P_{X_i,X_j}\}_{\{i,j\}\in\mathcal{E}}$. When $M \geq 3$, this algebra is strictly coarser than the algebra generated by $p(x_V)$. Marginal laws estimated on different realisations do not determine a common tuple, which is the observation model of ImageBind. Text and audio then meet only through separate image pairs.

Assumption 2 (Edge design)**.** *The edge set $\mathcal{E}$ is chosen independently of the latent variables. Conditional on $\{i, j\} \in \mathcal{E}$, the n_{ij} pairs are independent draws from the true joint P_{X_i,X_j}, and samples on distinct edges are independent.*

Assumption 2 is the case of edges missing completely at random. If pairing depends on the observed view, or on a factor that was never recorded, the observed conditional changes. Appendix L treats both cases. A missing entry still belongs to a sample that carries the other views, whereas a missing edge means that no common realisation was collected. The theorems below are statements about $\mathcal{F}_G$.

Definition 1 (Consistent block identifiability)**.** *Following Von Kügelgen et al. (2021); Daunhawer et al. (2023), a map g block identifies a factor c if $g(X) = h(c)$ almost surely for some smooth invertible h. A family $\{g_m\}$ identifies c consistently when a single map h works for every modality.*

Separate recovery on each modality is not enough for retrieval. Each encoder may apply its own diffeomorphism of c, and a query written in one chart then fails to match the same factor in another chart. Consistency requires one chart for the whole family. On an edge $\{i, j\}$ the population objective that expresses this requirement is

$$\mathcal{L}_{ij}(g_i, g_j) \;=\; \mathbb{E}\big[\|g_i(X_i) - g_j(X_j)\|^2\big] \;-\; H\big(g_i(X_i)\big) \;-\; H\big(g_j(X_j)\big), \tag{2}$$

where the expectation is under P_{X_i,X_j} and each encoder maps into the unit cube $(0, 1)^{d_s}$. The squared term forces the encoders to agree, and the entropy terms force each of them to use the cube. In the limit of infinitely many negatives and vanishing temperature, InfoNCE agrees with this objective (Oord et al., 2018; Zimmermann et al., 2021). The graph objective $\mathcal{L}_G = \sum_{\{i,j\}\in\mathcal{E}} w_{ij}\mathcal{L}_{ij}$ uses positive weights. Each term depends on one pairwise law, so $\mathcal{L}_G$ is a functional of $\mathcal{F}_G$.

4 Pairing graph identifiability

These edge losses identify one common chart only when the pairing graph lets the local charts be glued. A simultaneous joint would expose a larger algebra, but that joint is not observed. The argument therefore moves from population connectivity to induced subgraphs, then to the finite sample resistance of a hub, and finally to private factors and unpaired maps.

4.1 Population connectivity

Lemma 1 (Edge reduction, Daunhawer et al., 2023)**.** *Under Assumption 1, every population minimiser of $\mathcal{L}_{ij}$ satisfies $g_i(X_i) = g_j(X_j) = h_{ij}(z_s)$ almost surely, for a diffeomorphism h_{ij}, whenever the encoder dimension equals d_s and the edge carries no extra pairwise block.*

Lemma 1 is the two view input. Alignment, sufficient variability, and entropy on the cube produce a diffeomorphism of z_s, and Appendix C records the chain. If a pairwise block $z_{\{i,j\}}$ is present, the encoder dimension must match that block, as in Theorem 2.

Lemma 2 (Componentwise gauge group)**.** *Let $\{g_m\}$ minimise $\mathcal{L}_G$ in the population, and let C be a connected component of G. If ψ is a diffeomorphism of $(0, 1)^{d_s}$ that preserves volume, then the encoders $\tilde{g}_m = \psi \circ g_m$ for $m \in C$, and $\tilde{g}_m = g_m$ for every other modality, are still minimisers.*

Theorem 1 (Connectivity)**.** *Under Assumption 1, with encoder dimension d_s, let $\{g_m\}$ minimise $\mathcal{L}_G$ in the population. If G is connected, there is one diffeomorphism h such that $g_m(X_m) = h(z_s)$*

almost surely for every m, so z_s is consistently block identifiable. If G is disconnected, the gauge group of Lemma 2 acts separately on each component, and z_s is not consistently block identifiable.

Sufficiency is induction along a spanning tree. Lemma 1 makes each child agree with its parent, and entropy on the cube removes a leftover rotation, so the tree produces a single h (Zimmermann et al., 2021). Necessity is Lemma 2. A volume preserving change of coordinates on one component leaves every edge loss unchanged, and retrieval across the cut compares unrelated charts of z_s. The likelihood of Yao et al. (2024) depends on $p(x_V)$, so this group does not act on it. If the edges are whitened separately, the leftover lies in $\mathrm{GL}(d_s)$, and the relative maps come from a vertex potential if and only if G is connected, as recorded in Appendix D.3.

4.2 Induced subgraphs and stars

A block shared by only some of the modalities can be tied to a single chart only by pairs among those modalities. An edge whose two views do not both depend on the block leaves the gauge of that block free. Restricting Lemma 1 to the modalities in A, and applying Theorem 1 on the induced subgraph $G[A]$, gives the criterion for that block.

Theorem 2 (Induced subgraph)**.** *Under Assumption 1, a block z_A of dimension d_A is consistently block identifiable by encoders on A, trained on the edges $\mathcal{E} \cap \binom{A}{2}$, if and only if $G[A]$ is connected.*

Corollary 1 (Star)**.** *Let G be a star with hub h and spokes a, b. The block $z_{\{a,b\}}$, supported only on the two spokes, is not consistently block identifiable, whereas the triple intersection carried by $\{a, h, b\}$ is. The single edge $\{a, h\}$ identifies $S_a \cap S_h$, which may be a larger block. Recovering the global factor from a star is a subspace intersection computed at the hub.*

Proposition 1 (Hub intersection)**.** *In the linear Gaussian model of Section 4.3, the population canonical correlation subspace of the edge $\{h, m\}$, read at the hub, is $\mathrm{col}(A_{h \leftarrow C_{hm}})$ with $C_{hm} = S_h \cap S_m$. For a set R of spokes,*

$$\bigcap_{m \in R} \mathrm{col}(A_{h \leftarrow C_{hm}}) \;=\; \mathrm{col}\Big(A_{h \leftarrow \bigcap_{m \in R \cup \{h\}} S_m}\Big). \tag{3}$$

Retrieval between two spokes of a star returns the factors that both share with the hub, and it does not return sharing that never enters the hub. In the linear Gaussian specialisation, the cross covariances on the hub side are

$$\Sigma_{IT} \;=\; e_{I,s} e_{T,s}^\top + e_{I,IT} e_{T,IT}^\top, \qquad \Sigma_{IA} \;=\; e_{I,s} e_{A,s}^\top + e_{I,IA} e_{A,IA}^\top. \tag{4}$$

The column spaces at the hub intersect in $\mathrm{span}\{e_{I,s}\}$. The image and text covariance carries z_s together with the image and text block, and the image and audio covariance carries z_s together with the image and audio block, so the only direction present in both is the global factor. The loadings of z_{TA} appear in neither matrix, and no linear functional of the pair recovers a factor that neither edge observed. Spoke to spoke retrieval can therefore succeed for a class that passes through the hub and still fail for a factor that lives only on the two spokes. A jointly observed triple would expose that block through $p(x_I, x_T, x_A)$ (Yao et al., 2024), whereas independent star pairs do not determine this joint. The same split holds for edgewise InfoNCE once the encoder can carry both z_s and z_{TA}.

Encoder width is part of the criterion. At dimension d_s, a later text and audio edge cannot store z_{TA} without displacing z_s, while a wider encoder stores that block only after $G[\{T, A\}]$ is connected. The edge that recovers the block is the missing pair between the modalities that carry it, and an edge with vanishing shared signal does not, by Theorem 5.

4.3 Finite sample rates and hub design

The population theorems fix the chart when the pairwise laws are known exactly. With finitely many samples, each edge yields a noisy relative gauge, and the error after these gauges are glued is a property of G. The closed form below is for linear Gaussian mixers.

Assumption 3 (Linear Gaussian)**.** *The factors $z_s \sim \mathcal{N}(0, I_{d_s})$ and $z_{p,m} \sim \mathcal{N}(0, I_{d_m})$ are independent across blocks and modalities. The observation is $X_m = A_m z_s + B_m z_{p,m}$, where A_m has full column rank and $[A_m \; B_m]$ has a left inverse.*

On each edge the cross covariance is $\Sigma_{ij} = A_i A_j^\top$. A singular value decomposition of Σ_{ij} therefore recovers the column spaces $\mathrm{col}(A_i)$ and $\mathrm{col}(A_j)$. After whitening, alignment on that edge is a relative observation $\hat{r}_i - \hat{r}_j = \varepsilon_{ij}$ with $\mathrm{Cov}(\varepsilon_{ij}) \preceq (\sigma^2/n_{ij})I$. Let L be the Laplacian with weights $W_{ij} = n_{ij}$, and consider the least squares estimator of the vertex gauges in the translation gauge $\sum_m \hat{r}_m = 0$.

Theorem 3 (Spectral rate)**.** *Under Assumption 3, if G is connected, this estimator satisfies*

$$\mathbb{E}\|\hat{r}_u - \hat{r}_v\|^2 \;=\; \sigma^2\, R_{\mathrm{eff}}(u,v) \;\le\; \frac{2\sigma^2}{\lambda_2(L)}\, d_s \max_e n_e^{-1}, \tag{5}$$

together with $\mathbb{E}\|\hat{r} - r^\star\|_2^2 = \sigma^2 \mathrm{tr}(L^+) \le (M-1)\sigma^2/\lambda_2$. Equal sample sizes n on every edge give $L = nL_0$ and pairwise variance $\sigma^2/(n\lambda_2(L_0))$.

For this linear model, the Gauss and Markov theorem on the incidence matrix of G yields $\mathrm{Cov}(\hat{r}) = \sigma^2 L^+$ in the translation gauge. The risk

$$\mathbb{E}\|\hat{r} - r^\star\|_2^2 \;=\; \sigma^2\, \mathrm{tr}(L^+) \tag{6}$$

is therefore minimal among unbiased linear estimators of $r^\star$. Since $(e_u - e_v)^\top L^+ (e_u - e_v) = R_{\mathrm{eff}}(u,v)$ and Appendix B gives $R_{\mathrm{eff}} \le 2/\lambda_2$, the order $\Theta(1/\lambda_2)$ is tight for this noise. The same identity interpolates the population regimes. A bridge of very small weight makes λ_2 small, so the variance across it grows like $1/\lambda_2$, and Theorem 1 is the endpoint $\lambda_2 = 0$.

Remark 1 (Nonlinear mixers)**.** *The identity equation 6 is exact for Laplacian least squares under Gaussian edge noise. For a general mixer, or for InfoNCE with finitely many negatives, the excess between vertices u and v satisfies*

$$\mathrm{excess}(u,v) \;\le\; \frac{C}{n\lambda_2} + \frac{C_{\mathrm{degrade}}}{\lambda_2},$$

where $C_{\mathrm{degrade}} \ge 0$ depends on the mixer class and on the temperature, as recorded in Appendix M. If each edge identifies the shared block only up to an orthogonal factor, the same reduction produces the connection Laplacian of Bandeira et al. (2013), as recorded in Appendix E.4.

Theorem 4 (Star among trees)**.** *Among connected graphs on M vertices with $M-1$ edges, the star uniquely maximises λ_2, and the maximum value is 1. The path minimises λ_2, and the minimum value is $2\big(1-\cos(\pi/M)\big)$. Leaf to leaf resistance on the unit weight star is 2. End to end resistance on the unit weight path is $M-1$. These spectral values are recorded by De Abreu (2007). For identification, they say that a budget of $M-1$ datasets of equal size should be arranged as a hub when the goal is the worst pair variance in Theorem 3.*

With unit weights, a path multiplies the worst pair variance by $(M-1)/2$ relative to a hub, already 4.5 when $M = 10$. Further budget should join a spoke pair that Theorem 2 leaves unidentified, and the hub must also carry shared signal.

Theorem 5 (Collapse as a vanishing cut)**.** *For the scalar model $X_m = \alpha_m z_s + \sigma_m \varepsilon_m$, write $\eta_m = \alpha_m^2/\sigma_m^2$ for the ratio of shared signal to noise on modality m. Any encoder of X_m has mean squared error at least $1/(1+\eta_m)$ when z_s has unit variance. Two views have minimum mean squared error $1/(1+\eta_i+\eta_j)$. Weight each edge by $\rho_{ij}^2 n_{ij}$, where $\rho_{ij} = \sqrt{\frac{\eta_i}{1+\eta_i}\frac{\eta_j}{1+\eta_j}}$. If $\eta_h \to 0$ on a star, every edge weight tends to zero, and spoke to spoke error returns to the prior variance of z_s. A weak spoke removes only the edges that touch that spoke.*

A star is only as informative as its hub. If the hub carries no shared signal, every spoke weight vanishes and the weighted graph falls into isolated leaves, so modality collapse is a cut of that weighted graph. Finite sample canonical correlation analysis fails below the spike threshold for spiked covariance models (Perry et al., 2018), which is Theorem 3 in the regime $\sigma^2 \asymp \rho^{-2}$. A long path, or a hub with no shared signal, leaves the error floor of Theorem 5, and no encoder class undercuts that floor. Remark 2 records the same weights for neighbour pairs on an auxiliary graph.

Remark 2 (Effective pairing from a data graph)**.** *Let $H = (\mathcal{N}, \mathcal{F})$ be a graph on data indices whose nodes carry incomplete observations. If $\{u,v\} \in \mathcal{F}$ implies $z_s(u) = z_s(v) + \xi_{uv}$ with $\mathbb{E}\|\xi_{uv}\|^2 = \delta^2$, then a neighbour pair that carries modalities i and j is a noisy pairing edge in the sense of Theorem 8. The resulting effective graph G_{eff} has weights that grow with the number of such pairs and decay as $e^{-\delta^2}$. Under Assumptions 1 and 2, z_s is consistently block identifiable when G_{eff} is connected, with finite sample error tracking $\lambda_2(G_{\mathrm{eff}})$, by Theorems 1 and 8.*

4.4 PRIVATE FACTORS AND UNPAIRED MAPS

A private factor enters only one modality, so changing its coordinates leaves $\mathcal{L}_G$ unchanged. Anchors and the encoder dimension decide whether that factor is recovered or merely leaks, and a map of the marginals does not mark the shared coordinate.

Theorem 6 (Private factors)**.** *The group $\Psi = \prod_m \mathrm{Diff}(\mathcal{Z}_{p,m})$ acts by reparametrising the mixing maps and leaves every paired joint unchanged, so it leaves $\mathcal{L}_G$ unchanged. No functional of $\mathcal{L}_G$ separates $z_{p,m}$ from a diffeomorphic copy. If K anchors $a^{(k)} = \ell(z_{p,m}^{(k)})$ span fewer than d_m directions, a nontrivial stabiliser remains. At least d_m generic coordinate anchors reduce the problem to independent component analysis with an auxiliary variable (Hyvarinen et al., 2019; Khemakhem et al., 2020). A population content encoder of dimension d_s satisfies $I(g_m^\star(X_m); z_{p,m} \mid z_s) = 0$. Any invertible embedding $\iota_m(X_m)$ preserves the mutual information $I(X_m; z_{p,m})$.*

This invariance is the obstruction recorded by Hyvarinen et al. (2019), specialised to the paired joints that the loss sees. Fewer than d_m independent anchor directions leave a free reparametrisation, and an encoder wider than d_s can store private coordinates without identifying them.

Theorem 7 (Unpaired Brenier map and paired barycentre)**.** *Under Assumption 3, take $X_m = \alpha_m z_s e_1 + \sigma_m z_{p,m} e_2$. The quadratic Brenier map from P_{X_i} to P_{X_j} is $T(x) = \mathrm{diag}(\alpha_j/\alpha_i,\, \sigma_j/\sigma_i)\, x$. Its second coordinate is a rescaling of the private factor. The paired barycentre $\mathbb{E}[X_j \mid X_i]$ is a linear function of z_s.*

Unpaired optimal transport rescales the private coordinate together with the shared one, whereas the barycentre is a function of z_s. Diversified flow matching can identify a translation of diverse unpaired marginals, and that translation is distinct from $(z_s, z_{p,m})$ (Shrestha & Fu, 2025).

Proposition 2 (Flow matching inherits the coupling)**.** *Let $X_t = (1-t)X_0 + tX_1$ and $\dot{X}_t = X_1 - X_0$. The population flow matching velocity is $v_t^\star(x) = \mathbb{E}[\dot{X}_t \mid X_t = x]$ (Lipman et al., 2023; Tong et al., 2024). Under the independent coupling $P_{X_i} \otimes P_{X_j}$, the quadratic optimal deterministic map is the Brenier map of Theorem 7. Under the pairing coupling P_{X_i,X_j}, the same conditional expectation is the barycentre of that theorem. Optimal transport inside a minibatch of independently drawn endpoints converges to the Brenier map as the batch grows. A batch drawn from the true pairs follows the pairing coupling.*

Theorem 8 (Pairing noise)**.** *Suppose the shared factor on the side of modality j is observed as $z_s + \xi$ with $\mathbb{E}\|\xi\|^2 = \varepsilon^2$. Under Assumption 3, the excess mean squared error of predicting z_s from X_j, relative to the case $\varepsilon = 0$, equals*

$$\frac{\eta_j \varepsilon^2}{(1+\eta_j)(1+\eta_j+\eta_j\varepsilon^2)} \;=\; \Theta(\varepsilon^2) \tag{7}$$

as $\varepsilon \to 0$ with η_j held fixed. If the mixing maps are Lipschitz, the paired barycentre moves by $O(\varepsilon)$ in L^2. If they are only Hölder of order β, the movement is $O(\varepsilon^\beta)$.

A noisy pair behaves as a weaker edge, and in this Gaussian model the excess scales with the pairing noise variance. An independent redraw with probability γ multiplies the canonical correlation by $1-\gamma$. Pairing noise and a vanishing hub signal both reduce an edge weight in Theorem 3.

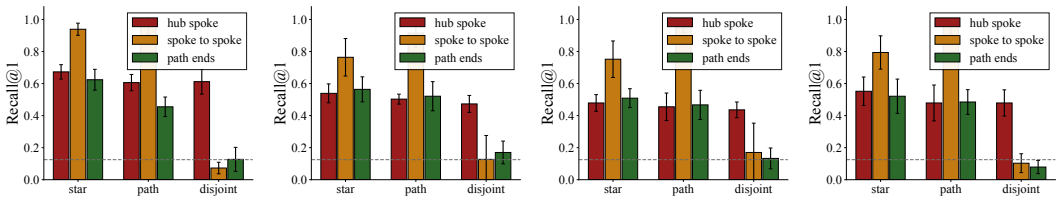
5 EXPERIMENTS

5.1 CONNECTIVITY ON TRAINED HEADS

Theorem 1 is a claim about the pairs a head is allowed to see. We therefore freeze the encoder and train a two layer Gaussian error linear unit projection head from random initialisation, by edgewise InfoNCE at temperature 0.07. The head observes only the edges of the current graph. CLIP ViT-B/32 supplies 512 dimensional image and templated text features, and Qwen3-VL at 2B, 4B, and 8B supplies vision and text states of dimension 2048, 2560, and 4096 (Bai et al., 2025). Each bank contains 100 Common Objects in Context images over eight classes, so Recall@1 chance is 0.125 (Lin et al., 2014). Five seeds are summarised by the mean and the standard deviation. A shuffle permutes pairs on every edge and leaves the frozen features unchanged. Edges draw independent

Table 1: Mean pair Recall@1 of trained CLIP and Qwen3-VL heads on a star, a path, disjoint spokes, and a pairing shuffle.

	star mean pair	path mean pair	disjoint spokes	shuffle mean pair
CLIP heads	0.772 ± 0.037	0.734 ± 0.021	0.073 ± 0.036	0.108
Qwen3-VL-2B	0.678 ± 0.046	0.713 ± 0.041	0.127 ± 0.149	0.165
Qwen3-VL-4B	0.646 ± 0.055	0.698 ± 0.060	0.170 ± 0.183	0.167
Qwen3-VL-8B	0.679 ± 0.083	0.717 ± 0.042	0.103 ± 0.059	0.192
chance	0.125	0.125	0.125	0.125



(a) CLIP heads. (b) Qwen3-VL-2B. (c) Qwen3-VL-4B. (d) Qwen3-VL-8B.

Figure 2: Edgewise InfoNCE on CLIP and Qwen3-VL heads. Red marks hub and spoke pairs, amber marks spoke to spoke pairs, green marks path ends, and the dashed line is chance 0.125.

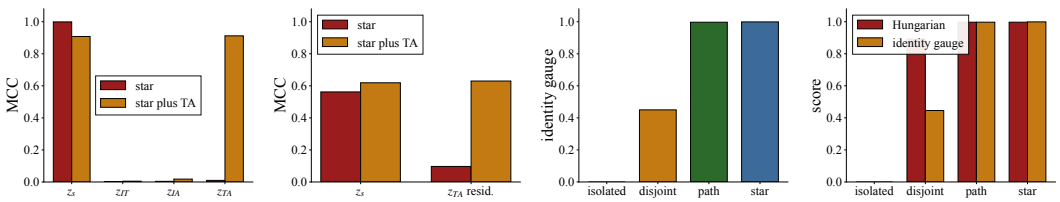
samples, and evaluation uses held out tuples that training never sees. Consistency across components is identity gauge correlation, defined in Appendix O. A post hoc canonical correlation probe on the same frozen banks is kept as a control and is never pooled with the trained numbers.

Connected graphs lie well above chance, and graphs that cannot share a chart do not. Table 1 reports the means, and Figure 2 breaks them out by edge type. Star and path mean pair Recall@1 lies between 0.65 and 0.77 on every encoder. Disjoint spokes fall to 0.073 ± 0.036 on CLIP, and on Qwen3-VL they stay compatible with chance once the standard deviation is included. The pairing shuffle matches those controls. Spoke to spoke class retrieval remains high on every connected graph, including the star. The views share the class factor, and the star still joins them through the hub, so Theorem 1 requires the alignment. Path end retrieval on CLIP is 0.624 for the star and 0.455 for the path under the same edge budget, which is the resistance gap of Theorems 3 and 4. Mean pair Recall on Qwen3-VL does not repeat that ranking, since an average over edges is dominated by easy pairs, and the disjoint standard deviations are large on only 33 evaluation tuples. A heterogeneous star, with a CLIP image hub and spokes in CLIP text and Qwen image features, still reaches spoke to spoke Recall@1 of 0.485 ± 0.131 against chance 0.125. The disconnected and shuffled controls sit at 0.212 ± 0.131 and 0.172 ± 0.014, as shown in Figure 4. The post hoc probe agrees, at 0.92 and 0.14 for CLIP star and disjoint spokes, and it does not replace training.

5.2 The block a star recovers

Corollary 1 identifies the triple intersection that passes through the hub, and it leaves unidentified any sharing that never enters the hub. The linear test uses orthogonal factors z_s, z_{IT}, z_{IA}, and z_{TA} in a three modality image hub world. Figure 3 shows the hub intersection. That one dimensional subspace matches z_s at mean correlation coefficient 0.999 and matches z_{TA} at 0.010. Canonical correlation analysis on an added text and audio edge recovers z_{TA} at 0.912. The two hub side covariances contain z_s together with the blocks that touch the hub. They do not contain the factor shared only by the spokes, so no linear functional of those covariances returns it. A jointly observed triple would have placed z_{TA} in the text and audio intersection. Pairwise star data do not.

The same factors, passed through tanh maps and trained by edgewise InfoNCE, repeat the split without a linear readout. On a star the heads recover z_s at mean correlation coefficient 0.56. After z_s is regressed out, residual correlation with z_{TA} stays at 0.061 and 0.134 on the two spokes. Adding the text and audio edge raises those residuals to 0.629 and 0.630, while the hub residual against z_{TA} stays near 0.08 on both graphs. The extra factor therefore appears only in the modalities that observe the new edge. Figure 3 places this residual beside the gauge check. An isolated vertex has correlation 0 with z_s, because no edge constrains it. A four modality path aligns its endpoints at 0.997 in the identity gauge. Two disjoint edges each recover z_s internally, with correlation near 0.96, yet identity



(a) Hub intersection. (b) InfoNCE residual. (c) Identity gauge. (d) Hungarian scores.

Figure 3: A star recovers the hub intersection and leaves an InfoNCE residual on z_{TA}. Identity gauge alignment and Hungarian scores on the same cut test whether the chart is shared.

gauge alignment across the cut is only 0.45. Hungarian matching on the same cut reaches 0.89, because it is free to realign each component on its own. That score is not a shared chart. Identity gauge correlation is the consistency statement in Theorem 1.

5.3 Resistance, private factors, and unpaired maps

Theorem 3 is an equality for linear Gaussian least squares, and only an upper bound with a nonnegative remainder for a nonlinear mixer. Four trees on ten vertices, each with nine edges, give Spearman correlation 1.00 between least squares error and both $\mathrm{tr}(L^+)$ and $1/\lambda_2$, and the Gaussian formula matches the measured error to within a few percent. Worst pair error grows with path length and stays flatter on a star, while path pairwise error tracks effective resistance with Spearman correlation 0.85. The same InfoNCE heads on a tanh class world give worst pair error about $1.191 + 0.090/\lambda_2$, with $R^2 = 0.929$ and a positive intercept. That intercept is the degrade term of Remark 1. Path length still changes the number of modalities together with λ_2. Holding $M = 8$ and the number of tree edges at seven, and varying only the tree, removes the confound. Worst pair error then tracks $1/\lambda_2$ with $R^2 = 0.997$ and Spearman correlation 1, and the star error 0.658 lies below the path error 1.959, as Theorem 4 requires. Path end Recall@1 on the CLIP heads follows the same ordering.

Contrastive losses see paired joints, so a shared encoder need not carry private coordinates, and an unpaired map need not mark which coordinate is shared. Two view canonical correlation analysis recovers shared factors at correlation 0.980 and private factors at 0.016. Linear anchors raise the private correlation from 0.45 with one anchor to above 0.99 once at least seven anchors are supplied. The count $d_p = 4$ is therefore enough only in population, and the extra anchors are a finite sample cost. On the diagonal Gaussian example the unpaired Brenier map reaches correlation 0.999 with both z_s and the private coordinate, because it rescales every direction it sees. The paired barycentre keeps correlation near 1 with z_s and lowers the private correlation to 0.52. Excess error under noisy pairing tracks ε^2 with Spearman correlation 1.00, in agreement with Equation equation 7. The minimum mean squared error formulas match to 1.4% and 1.7% for one view and for two views. Below the spiked covariance threshold, mean correlation over 25 trials stays between 0.04 and 0.08, then rises to 0.57 at spike 8, so a vanishing hub is this floor. Neighbour noise on a data graph drops canonical correlation from 0.997 to 0.024, in line with Remark 2. A 10 class tanh world repeats the connectivity gap, with star retrieval 1.00 against chance 0.10, path end retrieval 0.889, and cross component retrieval 0.035.

6 Conclusion

This paper gives an identifiability theory for multimodal training on independent pairwise collections. Modalities are the vertices of a pairing graph and observed pairs are its edges. A latent block is consistently block identifiable from the edgewise contrastive losses if and only if the modalities that carry it induce a connected subgraph. On a star the criterion recovers the hub intersection, and a class factor that passes through the hub is aligned across spokes. After each edge has estimated a local representation, residual encoder disagreement equals the effective resistance of the pairing network and scales as $1/\lambda_2$. Among trees the star uniquely minimises this residual when the hub carries the shared signal. The identity is an equality for linear Gaussian least squares and remains an upper bound, with a nonnegative remainder, for nonlinear mixers. Paired barycentres recover the shared factor, while a content encoder of that dimension keeps private coordinates out, and edgewise InfoNCE on CLIP and Qwen3-VL heads reproduces this ordering of graphs.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments, and assisting with translation. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, implementing methods, supporting qualitative and thematic data analysis, interpreting results. Besides, generating synthetic data sets, helping develop theoretical models or conceptual frameworks, cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for modifying scientific figures or images, editing a research paper to improve readability. We have reviewed all AI assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pp. 1247–1255. PMLR, 2013.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Afonso S Bandeira, Amit Singer, and Daniel A Spielman. A cheeger inequality for the graph connection laplacian. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1611–1630, 2013.

Yichao Cai, Yuhang Liu, Erdun Gao, Tianjiao Jiang, Zhen Zhang, Anton van den Hengel, and Prof Javen Qinfeng Shi. On the value of cross-modal misalignment in multimodal representation learning. *Advances in Neural Information Processing Systems*, 38:145859–145911, 2026.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, volume 26, 2013.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, pp. 1–23, 2023.

Nair Maria Maia De Abreu. Old and new results on algebraic connectivity of graphs. *Linear algebra and its applications*, 423(1):53–73, 2007.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pp. 217–227, 2020.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304, 2010.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in neural information processing systems*, volume 29, 2016.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217, 2020.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, pp. 1–33, 2023.

Yuhang Liu, Zhen Zhang, Dong Gong, Erdun Gao, Biwei Huang, Mingming Gong, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Beyond DAGs: A latent partial causal model for multimodal learning. In *International Conference on Learning Representations*, pp. 155872–155907, 2026.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international Conference on Machine Learning*, pp. 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359, 2020.

Yiwei Lou, Yuanpeng He, Rongchao Zhang, Yongzhi Cao, Hanpin Wang, and Yu Huang. Multitasks-based deep evidential fusion network for blind image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7512–7520, 2026.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and suboptimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.

Longfei Qiu, Jingqi Xiao, and Zhong Shao. Mechanized safety and liveness proofs for the mysticeti consensus protocol under the lido-dag framework. In *IEEE Symposium on Security and Privacy (SP)*, pp. 1842–1861. IEEE, 2026a.

Longfei Qiu, Jingqi Xiao, Ji-Yong Shin, and Zhong Shao. Suredistrib: Verifying almost-sure termination of composite asynchronous byzantine protocols. *Proceedings of the ACM on Programming Languages*, 10(PLDI):1152–1177, 2026b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Sagar Shrestha and Xiao Fu. Diversified flow matching with translation identifiability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 55360–55378, 2025.

Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011.

Yuanmin Tang, Jing Yu, Keke Gai, Gang Xiong, Gaopeng Gou, Meikang Qiu, and Qi Wu. Manipulation intention understanding for zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9466–9474, 2026.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with mini-batch optimal transport. In *Transactions on Machine Learning Research*, pp. 1–34, 2024.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939, 2020.

Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. *Advances in Neural Information Processing Systems*, 38:46634–46666, 2025.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, pp. 34817–34848, 2024.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In *International Conference on Learning Representations*, pp. 9588–9608, 2024.

Yinlin Zhu, Xunkai Li, Jishuo Jia, Miao Hu, Di Wu, and Meikang Qiu. Towards effective federated graph foundation model via mitigating knowledge entanglement. In *Advances in Neural Information Processing Systems*, volume 38, pp. 64599–64628, 2026.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International conference on machine learning*, pp. 12979–12990, 2021.

APPENDIX

A Notation and standing assumptions 14

B Preliminaries 14

C Edge reduction 15

D Connectivity and the gauge group 15

E Laplacian least squares 17

F Stars among trees 18

G Private factors 18

H Unpaired Brenier maps 19

I Pairing noise 20

J Collapse as a vanishing cut 20

K Data graph homophily 21

L MAR and NMAR pairing 21

M InfoNCE excess as a remark 21

N Additional spectral identities 22

O Experimental protocol and further results 24

A NOTATION AND STANDING ASSUMPTIONS

Logarithms are natural and are reported in nats. The pairing graph $G = ([M], \mathcal{E})$ is undirected and simple, and its weighted Laplacian is $L = D - W$ with $W_{ij} = W_{ji} \geq 0$, $W_{ii} = 0$, and $D_{ii} = \sum_j W_{ij}$. The associated quadratic form is $x^\top Lx = \sum_{\{i,j\}} W_{ij}(x_i - x_j)^2$. If G is connected then $\ker L = \mathrm{span}\{\mathbf{1}\}$, and the Moore and Penrose pseudoinverse L^+ inverts L on $\mathbf{1}^\perp$. The Fiedler value $\lambda_2(L)$ is the second smallest eigenvalue, and the effective resistance is $R_{\mathrm{eff}}(u, v) = (e_u - e_v)^\top L^+(e_u - e_v)$.

Table 2: Symbols used in the theorems.

symbol	meaning
$z_s, z_A, z_{p,m}$	global, subset, and private latent blocks
S_m	index set of blocks that enter modality m
f_m	mixing diffeomorphism of modality m
$G = ([M], \mathcal{E})$	pairing graph (observed pairwise joints)
n_{ij}	sample size on edge $\{i, j\}$
g_m	encoder of modality m
$\mathcal{L}_{ij}, \mathcal{L}_G$	edge and graph contrastive objectives
L, L^+, λ_2	Laplacian, pseudoinverse, Fiedler value
$R_{\mathrm{eff}}(u, v)$	effective resistance
$\eta_m = \alpha_m^2/\sigma_m^2$	shared SNR of modality m
ρ_{ij}	pairwise canonical correlation
G_{eff}	effective pairing graph induced by a data graph

Table 2 lists the symbols used in the theorems. Assumption 1 is the hypothesis for the population statements in Theorems 1, 2, and 6, while Assumption 3 is the hypothesis for the closed form rates, the Brenier maps, and the MMSE identities in Theorems 3, 7, 8, and 5. Assumption 2 in Section 3 treats the edge set as a design choice. The edge set is independent of the latents, and each observed edge is an independent draw from the true pairwise joint.

Assumption 4 (No extra pairwise block)**.** *The encoder dimension equals d_s, and $d_{\{i,j\}} = 0$ on every edge, or else the learner restricts attention to the globally shared block. Subset blocks whose encoder dimension matches the block are handled by Theorem 2. This hypothesis is the one used in Theorem 1.*

Remark 3 (Scope of the theorems)**.** *Private coordinates are recovered from $\mathcal{L}_G$ only when anchors are available, by Theorem 6. Unpaired Brenier maps mix z_s with private coordinates, by Theorem 7. The spectral facts in Theorem 4 are classical, and the identification content is the reduction of representation variance to λ_2. Hungarian matching is not a test of consistent identification, since it realigns independent gauges and hides the necessity half of Theorem 1.*

B PRELIMINARIES

Lemma 3 (Data processing)**.** *If $T = t(X)$ is measurable then $I(T; Z) \leq I(X; Z)$, with equality if T is sufficient for Z. If $X_m = f_m(z^{(m)})$ is a diffeomorphism then $\sigma(X_m) = \sigma(z^{(m)})$ and $I(X_m; z_s) = I(z^{(m)}; z_s)$.*

Lemma 4 (Alignment plus entropy)**.** *Let $g : \mathcal{X} \to (0, 1)^d$ be smooth and let P_X have a density that is strictly positive on a nonempty open set. Then $H(g(X))$ is maximised, among maps into the unit cube, if and only if $g(X)$ is uniform on $(0, 1)^d$. Uniformity plus smoothness and a nonvanishing density imply that g is a local diffeomorphism P_X-almost everywhere (Zimmermann et al., 2021). If additionally $\mathbb{E}\|g_i(X_i) - g_j(X_j)\|^2 = 0$, then $g_i(X_i) = g_j(X_j)$ almost surely.*

InfoNCE agrees with alignment plus entropy in the limit of infinitely many negatives and vanishing temperature (Oord et al., 2018; Gutmann & Hyvärinen, 2010; Zimmermann et al., 2021). The arguments below use equation 2 and do not rederive that limit.

Lemma 5 (Rayleigh and resistance)**.** *On a connected weighted graph,*

$$\lambda_2(L) \;=\; \min_{x \perp \mathbf{1},\, \|x\|_2 = 1} x^\top Lx.$$

For any vertices u, v, $R_{\text{eff}}(u, v) \le 2/\lambda_2(L)$, because $\|e_u - e_v\|_2^2 = 2$ and $e_u - e_v \perp \mathbf{1}$. On the unit weight star, the resistance between a leaf and the hub is 1 and the resistance between two leaves is 2. On the unit weight path, resistance equals graph distance.

Lemma 6 (Gaussian conditioning)**.** *If (U, V) is jointly Gaussian and centred, $\mathbb{E}[U \mid V] = \Sigma_{UV}\Sigma_{VV}^{-1}V$ and $\mathrm{Var}(U \mid V) = \Sigma_{UU} - \Sigma_{UV}\Sigma_{VV}^{-1}\Sigma_{VU}$.*

Lemma 7 (Classical Fiedler values)**.** *For the unweighted graphs on M vertices the following spectra are classical (De Abreu, 2007). The star $K_{1,M-1}$ has spectrum $\{0, 1^{M-2}, M\}$, so $\lambda_2 = 1$. The path P_M has $\lambda_k = 2\big(1 - \cos(\pi(k-1)/M)\big)$, so $\lambda_2 = 2(1 - \cos(\pi/M))$. The cycle C_M has $\lambda_2 = 2(1 - \cos(2\pi/M))$, and the complete graph K_M has $\lambda_2 = M$.*

C Edge reduction

Lemma 1 is the two view theorem of Daunhawer et al., specialised to a single pairing edge (Daunhawer et al., 2023). The chain is recorded so that later glueing steps do not cite that result as a black box.

Proof of Lemma 1. The squared alignment term is nonnegative and vanishes if and only if $g_i(X_i) = g_j(X_j) =: U$ almost surely. Hence U is measurable with respect to both $\sigma(X_i)$ and $\sigma(X_j)$. By Lemma 3 and Assumption 1, $U = u_i(z^{(i)}) = u_j(z^{(j)})$ almost surely. The coordinates common to $z^{(i)}$ and $z^{(j)}$ are $(z_s, z_{\{i,j\}})$, and the pairwise block is absent under Assumption 4. Fix those shared coordinates. Sufficient variability of the private factor of modality i, as required by Assumption 1, varies $z_{p,i}$ on a nonempty open set while the right-hand side u_j does not depend on $z_{p,i}$. Therefore u_i is constant in $z_{p,i}$. The same argument drops every factor supported on only one side of the edge. Thus $U = h(z_s)$ almost surely. Entropy maximisation on the cube, by Lemma 4, forces $h_{\#}p_{z_s}$ to be uniform on $(0, 1)^{d_s}$, hence h is a diffeomorphism.

If $d_{\{i,j\}} > 0$ and the encoder dimension is still d_s, alignment permits U to depend on $z_{\{i,j\}}$, but the cube cannot carry $(z_s, z_{\{i,j\}})$ invertibly. Entropy then selects a factor of dimension d_s with maximal mutual information, and the two blocks can be confused when they have equal dimension and equal variability. That is why Theorem 1 uses Assumption 4, and why Theorem 2 matches the encoder dimension to the target block.

If $z_s \not\perp z_{p,m}$, the same argument identifies the maximal common block, in the sense of von Kügelgen content, rather than a coordinate that has been defined independently of style. That convention is used throughout. □

D Connectivity and the gauge group

D.1 Sufficiency

Proof of Theorem 1, sufficiency. Assume G is connected and $\{g_m\}$ minimises $\mathcal{L}_G$ in the population. Lemma 1 supplies, on every edge $\{i, j\}$, a diffeomorphism h_{ij} with $g_i(X_i) = g_j(X_j) = h_{ij}(z_s)$ a.s. Fix a spanning tree T and a root 1. On every tree edge incident to the root, $g_j(X_j) = g_1(X_1)$, so h_{1j} coincides with the map h_1 defined by g_1. Induct along tree paths. Every vertex m satisfies $g_m(X_m) = g_1(X_1) = h_1(z_s)$ almost surely. Entropy regularisation pins every encoder to the same uniform cube, so there is no residual unknown rotation on edges. □

D.2 Necessity from independent gauges

Proof of Lemma 2. Let C be a connected component of G and ψ a volume-preserving diffeomorphism of $(0, 1)^{d_s}$. No edge of G has one end in C and the other outside C, so every alignment term of $\mathcal{L}_G$ is unchanged. Inside C, ψ preserves Lebesgue measure on the cube, hence entropy. The modified encoders therefore remain population minimisers. □

Proof of Theorem 1, necessity. Let C_1, C_2 be distinct connected components of G, and let $\{g_m\}$ be a population minimiser constructed as in the sufficiency argument (applied componentwise). The

group of volume-preserving diffeomorphisms of $(0,1)^{d_s}$ is nontrivial. Pick ψ in that group, $\psi \neq \mathrm{id}$, and set

$$\tilde{g}_m \;=\; \begin{cases} \psi \circ g_m & m \in C_1, \\ g_m & m \notin C_1. \end{cases}$$

There is no edge with one end in C_1 and the other outside C_1, so every alignment term is unchanged. Inside C_1, ψ preserves Lebesgue measure on the cube, hence $H(\tilde{g}_m) = H(g_m)$. Thus $\tilde{g}$ is still a minimiser. For $a \in C_1$ and $b \in C_2$, however, $\tilde{g}_a(X_a) = \psi(h(z_s))$ and $\tilde{g}_b(X_b) = h(z_s)$. No single diffeomorphism h' satisfies both identities unless $\psi = \mathrm{id}$. Hence z_s is not consistently block identified.

The extreme case $M = 2$ with $\mathcal{E} = \emptyset$ is two independent datasets. Any invertible encoding of each marginal into the uniform cube maximises entropy, there is no alignment term, and cross modal coordinates are undefined. □

D.3 $\mathrm{GL}(d_s)$ RESIDUAL AND SYNCHRONISATION

If each edge is whitened separately, Lemma 1 becomes $g_i(X_i) = Q_{ij} h(z_s)$ with $Q_{ij} \in \mathrm{GL}(d_s)$. Along a cycle $i_1 \to \cdots \to i_k \to i_1$ the product $Q_{i_1 i_2} \cdots Q_{i_k i_1}$ must be the identity for the assignment to be consistent. The resulting problem is $\mathrm{GL}(d_s)$ synchronisation on G. Relative observations $Q_{ij} = Q_i Q_j^{-1}$ are compatible if and only if they derive from a vertex potential $\{Q_m\}$. Connectivity of G is still necessary and sufficient for uniqueness of that potential up to a global $\mathrm{GL}(d_s)$ factor, which is removed by pinning $Q_{\mathrm{root}} = I$. Disconnected graphs yield one independent factor per component. Finite sample orthogonal residuals land on the connection Laplacian of Bandeira et al. (2013), as recorded in Appendix E.4.

D.4 INDUCED SUBGRAPHS

Proof of Theorem 2. Restrict the generative model to the modalities in A, treat every block not supported on A as private to those modalities, and apply Theorem 1 to the graph $G[A]$ with encoder dimension d_A. The sufficient-variability hypothesis of Assumption 1 is inherited by the complementary coordinates. Hence z_A is consistently block-identified by encoders on A trained on $\mathcal{E} \cap \binom{A}{2}$ if and only if $G[A]$ is connected. □

Proof of Corollary 1. On a star with hub h and spokes a, b, the induced graph $G[\{a, b\}]$ has no edge, so $z_{\{a,b\}}$ is not identifiable. The induced graph $G[\{a, h, b\}]$ is the path from a through h to b, which is connected, so the triple intersection carried by $\{a, h, b\}$ is identifiable. The single edge $\{a, h\}$ identifies the possibly larger intersection $S_a \cap S_h$. □

Proof of Proposition 1. Let $X_h = \sum_{B \ni h} A_{h,B} z_B + B_h z_{p,h}$ and likewise for spoke m. The cross-covariance of edge $\{h, m\}$ is

$$\Sigma_{hm} \;=\; \sum_{B \ni h,\, B \ni m} A_{h,B} A_{m,B}^{\top}.$$

If the column spaces of the blocks at h are in direct sum, which holds almost surely for Gaussian loadings and is enforced by orthogonal embeddings in the linear Gaussian checks of Appendix O, the column space of Σ_{hm} is the hub side CCA subspace $\mathrm{col}(\{A_{h,B} : B \ni h, m\}) = \mathrm{col}(A_{h \leftarrow C_{hm}})$ with $C_{hm} = S_h \cap S_m$. Intersecting over spokes $m \in R$ retains only those blocks B that contain every $m \in R$ and the hub, i.e. $B \supset R \cup \{h\}$. If the unique such block is the global z_s, the intersection is $\mathrm{col}(A_{h,s})$. □

The identifiability algebra of Yao et al. acts on jointly observed views (Yao et al., 2024). A jointly observed triple would expose $z_{\{a,b\}}$ as a selector on $p(x_a, x_h, x_b)$. Independent pairwise star samples never yield that joint, so the extra pairwise block remains invisible.

E LAPLACIAN LEAST SQUARES

E.1 EDGE NOISE FROM CCA

Under Assumption 3, $\Sigma_{ij} = A_i A_j^\top$ on edges. The SVD of Σ_{ij} recovers $\mathrm{col}(A_i)$ and $\mathrm{col}(A_j)$. With isotropic observation noise the canonical correlations equal

$$\rho_{ij} \;=\; \sqrt{\frac{\eta_i}{1+\eta_i}\cdot\frac{\eta_j}{1+\eta_j}}, \qquad \eta_m = \frac{\alpha_m^2}{\sigma_m^2}$$

in the scalar case, and a diagonal of such values in the vector case after simultaneous whitening.

Let $\hat{\Sigma}_{ij}$ be the empirical cross-covariance on n_{ij} pairs. Standard concentration gives $\|\hat{\Sigma}_{ij} - \Sigma_{ij}\|_{\mathrm{op}} = O_p(\sqrt{D_i D_j / n_{ij}})$. The Davis and Kahan theorem, in the form used for sample covariances (Yu et al., 2015), converts this into a subspace error of order $O_p(\|\hat{\Sigma}_{ij} - \Sigma_{ij}\|_{\mathrm{op}}/\mathrm{gap})$. The gap is $\Theta(1)$ without observation noise and $\Theta(\rho_{ij})$ with it. Denoting by $\hat{u}_m = \hat{v}_m^\top X_m$ the estimated canonical coordinate, the edge difference $\hat{u}_i - \hat{u}_j$ has variance $\le C\sigma^2/n_{ij}$ with $\sigma^2 \asymp \rho_{ij}^{-2}$. Vector valued shared factors repeat the scalar argument on each canonical axis, or replace Euclidean distance by a Procrustes residual, and the constants pick up a factor d_s.

E.2 LEAST SQUARES ON L

Let $r \in \mathbb{R}^{M\times d_s}$ have rows $r_m^\top$. The empirical glueing objective is $\sum_{\{i,j\}\in\mathcal{E}} W_{ij}\|r_i - r_j\|^2 = \mathrm{tr}(r^\top L r)$ with $W_{ij} = n_{ij}$, under the translation gauge $\mathbf{1}^\top r = 0$. The observation model on edges is $y_{ij} = r_i^\star - r_j^\star + \varepsilon_{ij}$ with $\mathbb{E}\varepsilon_{ij} = 0$ and $\mathrm{Cov}(\varepsilon_{ij}) \preceq (\sigma^2/n_{ij})I$.

Orient the edges arbitrarily and let B denote the (weighted) incidence matrix, so that $L = B^\top \mathrm{diag}(n_e) B$ when $W_{ij} = n_{ij}$. The centred least squares estimator satisfies $\hat{r} - r^\star = L^+ B^\top \varepsilon$ in the scalar case $d_s = 1$. Hence

$$\mathbb{E}\|\hat{r} - r^\star\|_2^2 = \sigma^2\,\mathrm{tr}\big(L^+ B^\top \mathrm{diag}(n_e^{-1}) B L^+\big). \tag{8}$$

A coarser bound uses $n_e^{-1} \le (\min_e n_e)^{-1}$ together with $\mathrm{tr}(L^+) \le (M-1)/\lambda_2$. A tighter and coordinatewise statement uses the resistance network whose edge conductances are the precisions n_{ij}/σ^2, namely

$$\mathbb{E}(\hat{r}_u - \hat{r}_v)^2 \;=\; \sigma^2\, R_{\mathrm{eff}}(u,v) \;\le\; \frac{2\sigma^2}{\lambda_2(L)}. \tag{9}$$

For $d_s > 1$ the coordinates are independent under Assumption 3 after whitening, and equation 9 is multiplied by d_s. This is equation 5 in the main text.

If every $n_{ij} = n$ then $L = nL_0$ for the unweighted Laplacian L_0, $\lambda_2(L) = n\lambda_2(L_0)$, $R_{\mathrm{eff}} = R_{\mathrm{eff}}^{(0)}/n$, and pairwise variance equals $\sigma^2/(n\lambda_2(L_0))$. The sample size needed to drive consistent identification error below ε therefore scales as λ_2^{-1}.

Proof of Theorem 3. The display equation 5 is equation 9 together with $R_{\mathrm{eff}} \le 2/\lambda_2$ from Lemma 5 and the d_s factor from independent coordinates. The global risk identity $\mathbb{E}\|\hat{r} - r^\star\|_2^2 = \sigma^2\mathrm{tr}(L^+)$ holds exactly when $\mathrm{Cov}(\varepsilon_{ij}) = (\sigma^2/n_{ij})I$ and $W_{ij} = n_{ij}$ in the scalar Gaussian model of Appendix E.3. The inequality $\mathrm{tr}(L^+) \le (M-1)/\lambda_2$ is spectral. □

E.3 GAUSS AND MARKOV TIGHTNESS

Lemma 8 (Gauss and Markov on the cut space)**.** *In the linear Gaussian edge noise model $y = Br^\star + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{diag}(n_e^{-1}))$, the least squares estimator $\hat{r} = L^+ B^\top \mathrm{diag}(n_e) y$, taken in the gauge $\mathbf{1}^\top r = 0$, is the minimum variance unbiased linear estimator of $r^\star$, and*

$$\mathrm{Cov}(\hat{r}) \;=\; \sigma^2 L^+$$

on $\mathbf{1}^\perp$. In particular $\mathbb{E}\|\hat{r} - r^\star\|_2^2 = \sigma^2\mathrm{tr}(L^+)$ cannot be improved by another unbiased linear estimator, so the order $\Theta(1/\lambda_2)$ is tight for this noise model.

Nonlinear mixings add an encoder complexity term. A standard Rademacher argument on a function class $\mathcal{G}$ with complexity $\mathfrak{R}_n$ yields excess risk $O(\mathfrak{R}_n + \sigma^2/(n\lambda_2))$ at a population minimiser. The exact identity $\sigma^2\mathrm{tr}(L^+)$ is for Laplacian least squares under Gaussian edge noise.

E.4 ORTHOGONAL RESIDUALS

If each edge identifies the shared block only up to an unknown rotation $Q_i Q_j^{-1} \in \mathrm{O}(d_s)$, the glueing problem is $O(d_s)$ synchronisation. Let $\hat{Q}_{ij}$ be the Procrustes alignment of the two CCA bases on edge $\{i, j\}$. The connection Laplacian L_1 of Bandeira et al. (2013) has a frustration constant η_G satisfying $\lambda_1(L_1) \leq \eta_G \leq C\lambda_1(L_1)/\lambda_2(L_0)$. We invoke their Cheeger inequality and do not re-prove it. The contribution is the reduction itself. Pairwise contrastive estimation produces exactly the relative observations that their spectral method consumes. Angular synchronisation in $d_s = 1$ is the eigenvector method of Singer (2011).

F STARS AMONG TREES

Proof of the star spectrum. Label the hub by 1 and the leaves by $2, \ldots, M$. Then $L\mathbf{1} = 0$. Any vector supported on two leaves with opposite signs, $x = (0, 1, -1, 0, \ldots)$, satisfies $Lx = x$, so 1 is an eigenvalue. The $(M-2)$-dimensional subspace of leaf vectors orthogonal to $\mathbf{1}$, equivalently the leaf coordinates that sum to zero with hub coordinate zero, is contained in that eigenspace. The remaining direction $x = (M-1, -1, \ldots, -1)$ satisfies $Lx = Mx$. Hence the spectrum is $\{0, 1^{M-2}, M\}$ and $\lambda_2 = 1$. □

Proof of the path spectrum. The eigenvectors of the path are discrete Dirichlet modes $x_j^{(k)} = \cos\big(\pi(k-1)(j-1/2)/M\big)$, with eigenvalues $2-2\cos(\pi(k-1)/M)$ (De Abreu, 2007). In particular $\lambda_2(P_M) = 2(1 - \cos(\pi/M))$. □

Lemma 9 (Trees have algebraic connectivity at most one)**.** *For every tree T on $M \geq 2$ vertices, $\lambda_2(T) \leq 1$, with equality if and only if T is a star (De Abreu, 2007).*

Proof. A tree has a leaf, so $\min_i \deg(i) = 1$. The bound $\lambda_2 \leq 1$ for trees follows from the variational characterisation together with a test function supported near a leaf, or from the diameter bound. A tree of diameter 2 is a star, and $\lambda_2 = 1$ there as above. If the diameter is at least 3, there is an induced path P_4. The path has $\lambda_2(P_4) = 2(1 - \cos(\pi/4)) = 2 - \sqrt{2} < 1$. Adding edges does not decrease λ_2 in the Loewner order, but here we compare trees, which are edge-minimal connected graphs. Among trees, diameter at least 3 forces $\lambda_2 < 1$ (De Abreu, 2007). Uniqueness of the maximiser is the uniqueness of diameter-2 trees.

The path minimises λ_2 among trees, because it maximises diameter, and the closed form $2(1 - \cos(\pi/M))$ is the standard minimum (De Abreu, 2007). The spectral comparison in Theorem 4 invokes Fiedler and Merris, and the identification content is the reduction of representation variance to λ_2. The numerical spectrum of the star, the path, the cycle, and the complete graph agrees with the closed forms to machine precision for $M \leq 15$, as reported in Appendix O.7. □

Proof of Theorem 4. Combine Lemma 7 with Lemma 9. For unit weight resistances, two star leaves form a series of two edges, so $R_{\mathrm{eff}} = 2$. Path vertices at distance k have $R_{\mathrm{eff}} = k$, so the worst pair has resistance $M - 1$. Substitute into Theorem 3. □

Adding an edge increases L in the Loewner order and cannot decrease λ_2. With budget $B > M-1$, the first extra edge on a star should join a spoke pair of largest resistance. Every spoke pair has resistance 2, and that is also the pair that Theorem 2 cannot see. Hub choice is not a purely graphical decision, because Theorem 5 requires the shared signal at the hub to stay away from zero.

G PRIVATE FACTORS

Proof of Theorem 6, group action. Let $\psi = (\psi_m) \in \Psi = \prod_m \mathrm{Diff}(\mathcal{Z}_{p,m})$. Reparametrise $z'_{p,m} = \psi_m(z_{p,m})$ and $f'_m(z_{\mathrm{sh}}, z'_{p,m}) = f_m(z_{\mathrm{sh}}, \psi_m^{-1}(z'_{p,m}))$. Then $X_m = f_m(z_{\mathrm{sh}}, z_{p,m}) = f'_m(z_{\mathrm{sh}}, z'_{p,m})$ as random variables, so every joint P_{X_i, X_j} is invariant. The objective $\mathcal{L}_G$ is a functional of those joints, hence invariant. No statistic of $\mathcal{L}_G$ can separate $z_{p,m}$ from $z'_{p,m}$. This is the obstruction recorded by Hyvarinen et al. (2019), specialised to the contrastive invariant. It is not a claim that nonlinear independent component analysis is impossible once auxiliary variables are present.

If the encoder dimension exceeds d_s, extra coordinates may store arbitrary functions of the private factors without increasing the alignment term, and entropy can even encourage packing them into the cube. That is overcomplete contamination, not identification. □

Proof of Theorem 6, anchors. An anchor $a^{(k)} = \ell(z_{p,m}^{(k)})$ is preserved by ψ_m if and only if $\ell \circ \psi_m = \ell$ on the observed points. If K differentials $D\ell(z_{p,m}^{(k)})$ span fewer than d_m directions, a nontrivial diffeomorphism remains in the stabiliser (move along the kernel, fix a complement). In the linear model $\ell(z) = Pz$ with $P \in \mathbb{R}^{d_a \times d_m}$, K samples give PZ. The latent z is identified only up to $\ker P$. If $d_a < d_m$ the kernel is perpetual. If $d_a \geq d_m$, P has full column rank, and $K \geq d_m$ generic samples, Z is invertible almost surely and the private block is recovered. The bound $K \geq d_m$ therefore assumes that ℓ is a coordinate embedding of sufficient ambient dimension. Sufficiency of many generic anchors is auxiliary variable independent component analysis (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Khemakhem et al., 2020). We reduce to that problem rather than reproving the identifiable variational autoencoder. The counting lower bound targets the contrastive stabiliser, not Locatello et al.'s pairs in which k factors change (Locatello et al., 2020). □

Proof of Theorem 6, leakage. A population content encoder of dimension d_s satisfies $g_m^\star(X_m) = h(z_s)$, so $\sigma(g_m^\star) \subseteq \sigma(z_s)$ and $I(g_m^\star; z_{p,m} \mid z_s) = 0$ by data processing. An invertible embedding $\iota_m(X_m)$ generates $\sigma(X_m) = \sigma(z^{(m)})$, hence $I(\iota_m; z_{p,m}) = I(X_m; z_{p,m})$. The chain rule gives $I(X_m; z_{p,m}) = I(z_s; z_{p,m}) + I(X_m; z_{p,m} \mid z_s)$. Under Assumption 3 the second term equals $H(z_{p,m})$, the Gaussian differential entropy of the private block. □

H Unpaired Brenier maps

Lemma 10 (Gaussian quadratic transport)**.** *The quadratic-cost optimal map from $\mathcal{N}(0, \Sigma_i)$ to $\mathcal{N}(0, \Sigma_j)$ is the linear map*

$$T(x) \;=\; \Sigma_i^{-1/2}\big(\Sigma_i^{1/2}\Sigma_j\Sigma_i^{1/2}\big)^{1/2}\Sigma_i^{-1/2}x$$

(Gabriel & Marco, 2019). If both covariances are diagonal, $T(x) = \mathrm{diag}(\Sigma_j^{1/2}\Sigma_i^{-1/2})x$ coordinatewise.

Proof of Theorem 7. Under the stated diagonal Gaussian model, $\Sigma_i = \mathrm{diag}(\alpha_i^2, \sigma_i^2)$ and $\Sigma_j = \mathrm{diag}(\alpha_j^2, \sigma_j^2)$. Lemma 10 yields $T(x) = \mathrm{diag}(\alpha_j/\alpha_i, \sigma_j/\sigma_i)\, x$. The second coordinate of $T(X_i)$ is $(\sigma_j/\sigma_i)X_i^{(2)} = \sigma_j z_{p,i}$, which is independent of z_s. Any representation that retains both coordinates of $T(X_i)$ is privately contaminated. Discarding the second coordinate requires already knowing which axis is shared, which is the identification problem. Unpaired optimal transport therefore does not label the shared axis.

The paired barycentre is Gaussian conditioning by Lemma 6,

$$\mathbb{E}[X_j \mid X_i] \;=\; \Sigma_{ji}\Sigma_{ii}^{-1}X_i \;=\; A_jA_i^\top(A_iA_i^\top + B_iB_i^\top)^{-1}X_i.$$

This is a contraction of X_i onto $\mathrm{col}(A_i)$ followed by the j-side shared loading, hence a (noisy) linear function of z_s. In the scalar diagonal model the coefficient of the shared coordinate is $\eta_i/(1+\eta_i)$, and $\mathrm{Corr}(\mathbb{E}[X_j \mid X_i], z_s) = \sqrt{\eta_i/(1+\eta_i)} > 0$ whenever $\alpha_i \neq 0$. □

Proposition 2 is stated in Section 4.4.

Proof of Proposition 2. The population velocity of linear-interpolant flow matching is the conditional expectation of $\dot{X}_t$ given X_t (Lipman et al., 2023; Tong et al., 2024). Under an independent coupling the endpoints are independent, so $\mathbb{E}[X_1 \mid X_0] = \mathbb{E}[X_1]$, which is 0 when centred, and the unique quadratic-optimal map of convex potential is Brenier's map, by Lemma 10. Under the pairing coupling the same conditional is the barycentre of Theorem 7. Mini batch optimal transport replaces the pairing coupling by an entropic plan computed from independently drawn pairs. As the batch grows, that plan converges to the marginal optimal transport plan, not to P_{X_i,X_j}. □

Diversified flow matching identifies a translation map under extra diversity assumptions (Shrestha & Fu, 2025). That object is not $(z_s, z_{p,m})$. The two identifiability statements can hold at the same time on unpaired and diverse marginals, while remaining distinct.

I PAIRING NOISE

Proof of Theorem 8, Gaussian. Let $X_i = \alpha_i z_s + \sigma_i \varepsilon_i$ and $X_j = \alpha_j(z_s + \xi) + \sigma_j \varepsilon_j$, with $z_s, \xi, \varepsilon_i, \varepsilon_j$ independent, $\mathrm{Var}(z_s) = 1$, $\mathrm{Var}(\xi) = \varepsilon^2$. Then $\mathrm{Cov}(X_i, X_j) = \alpha_i \alpha_j$ (the pairing noise ξ does not enter the cross-covariance if X_i is clean). Regression of z_s on X_j has

$$R^2_{z_s|X_j} \;=\; \frac{\alpha_j^2}{\alpha_j^2(1+\varepsilon^2) + \sigma_j^2} \;=\; \frac{\eta_j}{\eta_j(1+\varepsilon^2) + 1},$$

hence

$$\mathrm{MMSE}(z_s \mid X_j) \;=\; \frac{1 + \eta_j \varepsilon^2}{1 + \eta_j + \eta_j \varepsilon^2}.$$

At $\varepsilon = 0$ this is $1/(1+\eta_j)$. The excess over perfect pairing is

$$\frac{1 + \eta_j \varepsilon^2}{1 + \eta_j + \eta_j \varepsilon^2} - \frac{1}{1+\eta_j} \;=\; \frac{\eta_j \varepsilon^2}{(1+\eta_j)(1 + \eta_j + \eta_j \varepsilon^2)} \;=\; \Theta(\varepsilon^2)$$

as $\varepsilon \to 0$ with η_j fixed. The same expansion holds for the R^2 of X_j against X_i, which carries an extra factor $\eta_i/(1+\eta_i)$. If pairing is replaced by random mismatch (a pair is independently redrawn with probability γ), the canonical correlation is multiplied by $1 - \gamma$ and the excess is $\Theta(\gamma)$. □

Proof of Theorem 8, Lipschitz and Hölder. Let $X_j^\varepsilon = f_j(z_s + \xi, z_{p,j})$ and $X_j^0 = f_j(z_s, z_{p,j})$. If $\|f_j\|_{\mathrm{Lip}} \le L$ then $\mathbb{E}\|X_j^\varepsilon - X_j^0\|^2 \le L^2 \varepsilon^2$. The barycentre is an L^2 projection, so

$$\big\|\mathbb{E}[X_j^\varepsilon \mid X_i] - \mathbb{E}[X_j^0 \mid X_i]\big\|_{L^2} \;\le\; \|X_j^\varepsilon - X_j^0\|_{L^2} \;\le\; L\varepsilon.$$

Comparing with $h(z_s) = \mathbb{E}[X_j^0 \mid z_s]$ uses contractivity of conditional expectation.

If f_j is only Hölder-β, the same comparison yields $O(\varepsilon^\beta)$ in L^2. A one dimensional example is $f(z) = \mathrm{sign}(z)|z|^\beta$ on $z_s \sim \mathrm{Unif}[-1, 1]$. Interior increments scale as ε^β, and the barycentre cannot be smoother than f, as shown in Appendix N.6. □

J COLLAPSE AS A VANISHING CUT

Proof of Theorem 5. Scalar Gaussian $X_m = \alpha_m z_s + \sigma_m \varepsilon_m$ with unit-variance independent factors. Lemma 6 gives $\mathbb{E}[z_s \mid X_m] = \frac{\alpha_m}{\alpha_m^2 + \sigma_m^2} X_m$ and

$$\mathrm{Var}(z_s \mid X_m) \;=\; \frac{\sigma_m^2}{\alpha_m^2 + \sigma_m^2} \;=\; \frac{1}{1+\eta_m}.$$

Any measurable encoder of X_m has MSE at least this MMSE. Two views with independent observation noise add precisions, so $\mathrm{Var}(z_s \mid X_i, X_j)^{-1} = 1 + \eta_i + \eta_j$, hence $\mathrm{MMSE} = 1/(1+\eta_i+\eta_j)$. The canonical correlation is $\rho_{ij} = \sqrt{\frac{\eta_i}{1+\eta_i}\frac{\eta_j}{1+\eta_j}}$.

Weighting pairing edges by $W_{ij} = \rho_{ij}^2 n_{ij}$ produces a Laplacian to which Theorem 3 applies. On a star, $\eta_h \to 0$ sends every $\rho_{hm} \to 0$, hence every weight, hence $L \to 0$ and $\lambda_2 \to 0$. Identification between spokes has no direct edge. Unpaired samples from distinct realisations have product marginals, so $I(X_a; X_b) = 0$, and the estimator cannot beat the worse single view MMSE. A single weak spoke zeroes only the edges incident to that spoke, while the remaining spokes still communicate through a healthy hub. □

Finite sample CCA on a spiked cross covariance fails when $\rho^2 n/D$ drops below the spike threshold for spiked covariance models, which is of order one (Perry et al., 2018). We treat that failure as an operational extreme of Theorem 3 with $\sigma^2 \asymp \rho^{-2}$, not as a contribution to random matrix theory. Appendix O reports an empirical curve of mean correlation against the spike. The MMSE identities hold to relative error 1.4% for one view and 1.7% for two views.

K DATA GRAPH HOMOPHILY

Remark 2 is stated in Section 4.3.

Proof of Remark 2. Each homophilous neighbour pair is a pairing sample of noise level δ, in the sense of Theorem 8. The resulting pairwise datasets define G_{eff}. Theorem 1 applies to G_{eff} in the population limit $\delta = 0$ with infinitely many effective pairs on each edge of G_{eff}. Theorems 3 and 8 cover finite samples and $\delta > 0$. If H has Cheeger constant $h(H)$ and each node observes a uniformly random modality, neighbour pairs cover every modality pair at rate $\Theta(h(H))$. This is a combinatorial condition on G_{eff}. It is not a theorem about the depth of a graph neural network, about oversmoothing, or about the expressivity of message passing. □

L MAR AND NMAR PAIRING

Let $R_{ij} \in \{0, 1\}$ denote the indicator that a candidate pair is observed. If the pair is missing at random, in the sense that $P(R_{ij} = 1 \mid X_i, X_j) = P(R_{ij} = 1 \mid X_i)$, then $P(X_j \mid X_i, R_{ij} = 1) = P(X_j \mid X_i)$. The alignment term of $\mathcal{L}_{ij}$ depends on the paired joint only through that conditional, so it still targets the true conditional expectation. Lemma 1 is unchanged, provided the conditional still satisfies sufficient variability of the private factors.

If instead R_{ij} depends on unobserved private factors of modality j, for instance when captions are written only for objects that are easy to name, the observed paired joint is a selected version of P_{X_i,X_j}. The identified block then shrinks to the content that is invariant under that selection, in the sense of Cai et al. (2026). Theorem 1 is stated under the edge design of Assumption 2, and the object under this selection is that smaller invariant.

A graph that is connected but has a bridge of vanishing weight has small λ_2. Theorem 3 already gives the corresponding quantitative statement, since identification variance on either side of the bridge diverges as $1/\lambda_2$. The binary split between connected and disconnected graphs is the endpoint $\lambda_2 = 0$ of that continuum. If Assumption 1 fails for a frozen private coordinate of modality m, Lemma 1 permits the encoder to depend on that coordinate, and the identified object becomes a larger content block that includes the frozen style, as discussed by Daunhawer et al. (2023). In the language of Theorem 2, the frozen coordinate is absorbed into some z_A, and the induced subgraph criterion applies to the enlarged block.

M INFONCE EXCESS AS A REMARK

This section does not introduce a numbered theorem. It records conditions under which a finite negative InfoNCE minimiser can be fed into the resistance glue of Theorem 3.

Let $(\hat{g}_i, \hat{g}_j)$ be an ε minimiser of the empirical InfoNCE loss on edge $\{i, j\}$. The population minimiser agrees with $h(z_s)$ under Assumption 1 and Lemma 1, and three conditions turn the excess ε into an alignment residual. First, each g_m is Lipschitz and lands on the unit sphere, with temperature τ held fixed. InfoNCE is then Lipschitz in the encoders, and a Rademacher argument supplies $\varepsilon_n = O(\mathfrak{R}_n + \sqrt{\log(1/\delta)/n})$. Second, negatives drawn inside the batch must not reduce the loss to a linear readout of the class. A check that can be verified in practice is that a pairing shuffle and a disconnected graph both stop at chance, which is the control used in Section 5. When that check fails, the excess may be a class centre rather than an instance alignment. Third, the encoder dimension must match the target block. Dimension d_s can carry only z_s, and dimension at least $d_s + d_{TA}$ is required before z_{TA} can occupy leftover coordinates after a text and audio edge is added.

Under those conditions,

$$\mathbb{E}\|\hat{g}_i - \hat{g}_j\|^2 \;\le\; C_\tau \varepsilon + C_{\text{neg}} + C_{\text{degrade}},$$

and the resistance glue of Theorem 3 yields

$$\text{excess}(u, v) \;\le\; \frac{C}{n\lambda_2} + \frac{C_{\text{degrade}} + C_{\text{neg}}}{\lambda_2}.$$

The identity $\mathrm{MSE} = \sigma^2 R_{\mathrm{eff}}/n$ remains reserved for Assumption 3. The experiments in Section 5 support a positive slope in $1/\lambda_2$ on nonlinear worlds and do not support a tight $\Theta(1/(n\lambda_2))$ rate for InfoNCE.

N Additional spectral identities

N.1 Quadratic form, cut, and Cheeger

For any $S \subset [M]$, the weighted cut is $\mathrm{cut}(S) = \sum_{i \in S, j \notin S} W_{ij}$ and

$$\mathbf{1}_S^\top L \mathbf{1}_S \;=\; \mathrm{cut}(S).$$

Cheeger's inequality relates λ_2 to the bottleneck ratio $h_G = \min_S \mathrm{cut}(S)/\min(\mathrm{vol}(S), \mathrm{vol}(S^c))$. The Cheeger constant is not used as a substitute for λ_2. Theorem 3 is stated in λ_2 and R_{eff}, which are the exact constants of the least squares problem. The cut language is used only in Theorem 5, where a hub with $\eta_h \to 0$ zeroes every star edge and the weighted cut between any spoke and the rest vanishes.

N.2 Pseudoinverse on $\mathbf{1}^\perp$

Let $L = \sum_{k=2}^{M} \lambda_k u_k u_k^\top$ be the spectral decomposition with $\lambda_1 = 0$, $u_1 = \mathbf{1}/\sqrt{M}$, and $\lambda_2 = \lambda_2$. Then $L^+ = \sum_{k=2}^{M} \lambda_k^{-1} u_k u_k^\top$ and

$$\mathrm{tr}(L^+) \;=\; \sum_{k=2}^{M} \lambda_k^{-1} \;\le\; \frac{M-1}{\lambda_2}.$$

Equality holds if and only if $\lambda_2 = \cdots = \lambda_M$, which means that the graph is complete up to a weight. On a star, $\mathrm{tr}(L^+) = (M-2)\cdot 1 + 1/M = (M-1)^2/M$, because the eigenvalue 1 has multiplicity $M-2$ and the remaining eigenvalue is M. On a path the sum $\sum_{k=2}^{M} \big(2 - 2\cos(\pi(k-1)/M)\big)^{-1}$ has no simpler closed form. Numerically $\mathrm{tr}(L^+(P_{10})) = 16.5$, as reported in Table 7.

N.3 Effective resistance as a voltage

Supplying unit current at u and extracting it at v yields voltages ϕ solving $L\phi = e_u - e_v$ in the gauge $\mathbf{1}^\top \phi = 0$, so $\phi = L^+(e_u - e_v)$, and the dissipated energy is $\phi_u - \phi_v = R_{\mathrm{eff}}(u, v)$. Series and parallel laws recover the star and path formulae used in Theorem 4 without invoking the spectral definition.

N.4 Explicit star eigenvectors

For the unit weight star, one orthonormal eigenbasis takes $u_1 = \mathbf{1}/\sqrt{M}$ and

$$u_M = \big(\sqrt{(M-1)/M},\, -1/\sqrt{M(M-1)}, \ldots,\, -1/\sqrt{M(M-1)}\big)$$

for eigenvalue M, together with any orthonormal basis of $\{x : x_1 = 0, \sum_{m \ge 2} x_m = 0\}$ for eigenvalue 1. The resistance between two leaves is

$$R_{\mathrm{eff}}(2,3) \;=\; (e_2 - e_3)^\top L^+ (e_2 - e_3) \;=\; 1^{-1} \|P_{\lambda=1}(e_2 - e_3)\|_2^2 \;=\; 2,$$

since $e_2 - e_3$ already lives in the $\lambda = 1$ subspace.

N.5 Cycle and complete graph

The cycle C_M is the circulant Laplacian with first row $(2, -1, 0, \ldots, 0, -1)$ and eigenvalues $2 - 2\cos(2\pi k/M)$ for $k = 0, \ldots, M-1$, so $\lambda_2 = 2 - 2\cos(2\pi/M) = 4\sin^2(\pi/M)$. The complete graph is $L = MI - \mathbf{1}\mathbf{1}^\top$, so every vector orthogonal to $\mathbf{1}$ is an eigenvector with eigenvalue M. The worst pair resistance on K_M is $2/M$. These identities are checked numerically in Table 3.

The maximum absolute error over $M = 3, \ldots, 15$ is below 10^{-14}. Resistance between two leaves of the star is 2, and resistance between the two ends of the path is $M - 1$.

Table 3: Numerical and closed form Fiedler values.

M	star	path $2(1-\cos\pi/M)$	cycle $2(1-\cos 2\pi/M)$	K_M
3	1	1	3	3
4	1	0.5858	2	4
5	1	0.3820	1.3820	5
6	1	0.2679	1	6
7	1	0.1981	0.7530	7
8	1	0.1522	0.5858	8
9	1	0.1206	0.4679	9
10	1	0.0979	0.3820	10
11	1	0.0810	0.3175	11
12	1	0.0681	0.2679	12
13	1	0.0581	0.2291	13
14	1	0.0501	0.1981	14
15	1	0.0437	0.1729	15

N.6 Hölder pairing noise, one dimension

Let $z_s \sim \mathrm{Unif}[-1,1]$, $\xi \sim \mathrm{Unif}[-\varepsilon,\varepsilon]$ independent of z_s, and $f(z) = \mathrm{sign}(z)|z|^\beta$ for $\beta \in (0,1]$. On the event $\{|z_s| > 2\varepsilon\}$,

$$\big|f(z_s+\xi) - f(z_s)\big| \;\le\; C_\beta\,|\xi|^\beta \;\le\; C_\beta \varepsilon^\beta,$$

because f is Hölder-β on $\mathbb{R}\setminus\{0\}$ with locally bounded constant away from the origin. Near the origin, $|f| \le \varepsilon^\beta$ on $[-\varepsilon,\varepsilon]$, so the contribution of that event to L^2 is of order $\varepsilon\cdot\varepsilon^{2\beta}$ in probability mass times height, which is $o(\varepsilon^{2\beta})$ as $\varepsilon\to 0$. Thus $\mathbb{E}|f(z_s+\xi)-f(z_s)|^2 = \Theta(\varepsilon^{2\beta})$. Conditional expectation onto X_i, taking $X_i = z_s$ observed without noise in this example, cannot reduce that order, and the barycentre $\mathbb{E}[f(z_s+\xi)\mid z_s]$ differs from $f(z_s)$ by $\Theta(\varepsilon^\beta)$ in L^2. This matches the Hölder clause of Theorem 8. The Gaussian and Lipschitz clauses are the ones used in the experiments. A matching lower bound for arbitrary Hölder mixings in higher dimension is left open.

N.7 Two-view Gaussian MMSE algebra

Let $X_i = \alpha_i z_s + \sigma_i\varepsilon_i$, $X_j = \alpha_j z_s + \sigma_j\varepsilon_j$, all standard normals independent. The stacked observation $X = (X_i, X_j)$ has

$$\Sigma_{Xz_s} \;=\; \begin{pmatrix}\alpha_i\\ \alpha_j\end{pmatrix}, \qquad \Sigma_{XX} \;=\; \begin{pmatrix}\alpha_i^2+\sigma_i^2 & \alpha_i\alpha_j\\ \alpha_i\alpha_j & \alpha_j^2+\sigma_j^2\end{pmatrix}.$$

The precision of z_s given X is

$$1 + \Sigma_{Xz_s}^\top \Sigma_{XX}^{-1}\Sigma_{Xz_s} \;=\; 1+\eta_i+\eta_j,$$

which is most easily seen in the information form, since each channel contributes an independent Gaussian likelihood of precision η_m. Hence $\mathrm{Var}(z_s \mid X_i, X_j) = 1/(1+\eta_i+\eta_j)$. The canonical correlation is the cosine of the angle between the two scalar observations after whitening,

$$\rho_{ij} \;=\; \frac{\alpha_i\alpha_j}{\sqrt{(\alpha_i^2+\sigma_i^2)(\alpha_j^2+\sigma_j^2)}} \;=\; \sqrt{\frac{\eta_i}{1+\eta_i}\frac{\eta_j}{1+\eta_j}}.$$

In particular $\rho_{ij} = 0$ if and only if $\alpha_i\alpha_j = 0$, so at least one view carries no shared signal. A hub with $\eta_h = 0$ on a star therefore contributes $\rho_{hm} = 0$ on every spoke, which is the cut-vanishing statement of Theorem 5.

If the j-view is corrupted by pairing noise $z_s+\xi$ as in Appendix I, the information form becomes

$$\mathrm{Var}(z_s \mid X_j)^{-1} \;=\; 1 + \frac{\eta_j}{1+\eta_j\varepsilon^2},$$

recovering $\mathrm{MMSE} = (1+\eta_j\varepsilon^2)/(1+\eta_j+\eta_j\varepsilon^2)$.

N.8 Gaussian Brenier algebra

For centred Gaussians the 2-Wasserstein distance has the closed form

$$W_2^2(\mathcal{N}(0,\Sigma_i),\mathcal{N}(0,\Sigma_j)) \;=\; \mathrm{tr}\big(\Sigma_i+\Sigma_j-2(\Sigma_i^{1/2}\Sigma_j\Sigma_i^{1/2})^{1/2}\big)$$

The map of Lemma 10 is the unique gradient of a convex potential pushing $\mathcal{N}(0, \Sigma_i)$ to $\mathcal{N}(0, \Sigma_j)$ (Gabriel & Marco, 2019). In the diagonal example of Theorem 7 both the shared and the private axes are rescaled, and W_2 does not know which axis is content.

A non-diagonal joint Gaussian mixing $X_m = A_m z_s + B_m z_{p,m}$ yields covariances $\Sigma_m = A_m A_m^\top + B_m B_m^\top$. The Brenier map is the symmetric positive definite alignment of Σ_i with Σ_j and therefore mixes $\mathrm{col}(A_i)$ with $\mathrm{col}(B_i)$ whenever those subspaces are not already the eigenbasis of both covariances. For generic loadings they are not. The paired regression $\mathbb{E}[X_j \mid X_i]$ always projects onto $\mathrm{col}(A_i)$ first, because $\Sigma_{ji} = A_j A_i^\top$ has image in $\mathrm{col}(A_j)$ and kernel containing $\mathrm{col}(B_i)^\perp$ after the Σ_{ii}^{-1} whitening.

Rectified flow (Liu et al., 2023) and flow matching under an independent coupling (Lipman et al., 2023) learn velocities of an interpolant whose law at the endpoints is the pair of marginals. Without a pairing coupling, the interpolant does not see P_{X_i, X_j}. Mini batch optimal transport (Tong et al., 2024; Cuturi, 2013) inserts an entropic coupling inside each batch. As the batch size tends to infinity that coupling tends to the marginal optimal transport plan, which is the Brenier map in the Gaussian quadratic case, and not to the pairing plan.

O EXPERIMENT PROTOCOL

This appendix records the generative worlds, the estimators, and the frozen embedding protocol used in Section 5. Reported point estimates use a fixed random seed. Recall@1 on frozen embeddings is averaged over three independent pairing splits.

O.1 GENERATIVE WORLDS

In the linear Gaussian world, $z_s \sim \mathcal{N}(0, I_{d_s})$ and $z_{p,m} \sim \mathcal{N}(0, I_{d_p})$ are independent, and $X_m = A_m z_s + B_m z_{p,m} + \sigma \varepsilon_m$. Loadings are drawn Gaussian and orthogonalised by a QR factorisation so that the shared and private column spaces are in direct sum, as required by Proposition 1 and by Table 6. The ImageBind toy uses three modalities I, T, A, four scalar factors $z_s, z_{IT}, z_{IA}, z_{TA}$, and three private factors. Modality I sees $\{z_s, z_{IT}, z_{IA}, z_{p,I}\}$, and the other modalities are filled in the same cyclic way. The pairing edges are $\{I, T\}$ and $\{I, A\}$, with $\{T, A\}$ added only when that comparison is required. Each edge draws its own n pairs, and realisations are not shared across edges.

The diagonal Brenier world has $D_m = 2$ and $X_m = (\alpha_m z_s, \sigma_m z_{p,m})$ with $\alpha_m, \sigma_m > 0$. The nonlinear class conditional world uses four modalities that share a 10 class one hot factor, mixed by fixed tanh maps that differ by modality, together with private Gaussian noise. Each pairing edge draws an independent batch, and evaluation uses held out complete tuples. Edgewise CCA of rank $k = 8$ is glued by orthogonal Procrustes on the training observations of the shared vertex, and is then applied to centred evaluation views. Recall@1 is the class match of the nearest neighbour in the glued embedding. A two view encoder with one hidden layer of rectified linear units, of width 32, is trained by alignment plus variance on the same world, with a shuffled pairing control. Shared encoder InfoNCE trains one such encoder per modality, and each step draws a minibatch from a single pairing edge, which is the ImageBind observation model. Spoke to spoke Recall@1 on the star is 1.00, and cross component Recall@1 on the disjoint graph is 0.018.

The nonlinear sign check uses $X_m = \tanh(A_m z_s + B_m z_{p,m})$. Random Fourier features of dimension 200, followed by linear CCA, provide a two view sign check of Theorem 1.

O.2 ESTIMATORS

Population and empirical CCA are the singular value decomposition of the cross covariance. Graph least squares sign aligns the one dimensional CCA coordinates on each edge and then solves $Lr = b$ in the gauge $\sum_m r_m = 0$. The unpaired Brenier map is the closed form Gaussian map computed from the empirical covariances, and the paired barycentre is ridge regression of X_j on X_i. Mean correlation after Hungarian matching of coordinates is the score of Von Kügelgen et al. (2021). Identity gauge correlation uses no further permutation, and it is the score whenever consistency

across components is the claim. Gaussian mutual information is $\frac{1}{2}\log\big(|\Sigma_U||\Sigma_V|/|\Sigma_{UV}|\big)$, which is exact on the linear Gaussian worlds.

O.3 Frozen multimodal embeddings

CLIP ViT-B/32 image embeddings and three independently templated text embeddings of the same COCO image are treated as four modalities. The text views use short label prompts of the form *a photo of a* $\{\cdot\}$, attribute prompts of the form *a* $\{\cdot\}$ *in natural light*, and caption prompts of the form *a photograph showing a* $\{\cdot\}$, each sampled from a bank of three templates. Qwen3-VL at 2B, 4B, and 8B uses mean pooled vision tokens for the image view and the last token hidden state for the three text views, with hidden widths 2048, 2560, and 4096. Class labels are nearest neighbour assignments in CLIP on eight COCO categories with at least six images each, and the pool has size $n = 100$. Pairing edges are independent resamples of the training pool, and evaluation is Recall@1 on a held out set of complete tuples with $n_{\text{eval}} = 33$, averaged over three seeds. Table 4 reports label and attribute retrieval on views that were never paired, which is the ImageBind style comparison of spokes. The table is a probe applied after the embeddings are frozen, and it is not a training experiment.

Table 4: Recall@1 between a label view and an attribute view that were never paired. Chance is 0.125. The probe is applied after the embeddings are frozen, and the training experiment is Table 1. Entries are mean $\pm$ standard deviation over three seeds.

embedding	star	two disjoint edges
CLIP ViT-B/32	0.919 ± 0.094	0.141 ± 0.200
Qwen3-VL-2B	0.616 ± 0.076	0.101 ± 0.014
Qwen3-VL-4B	0.455 ± 0.108	0.131 ± 0.100
Qwen3-VL-8B	0.616 ± 0.029	0.121 ± 0.025

O.4 Trained projection heads

The main text trains a two layer GELU head on each modality by cycling InfoNCE over the edges of the current pairing graph. Encoders remain frozen. Evaluation is class Recall@1 on held out complete tuples ($n_{\text{eval}} = 33$, five seeds). Table 1 is the CLIP and Qwen3-VL 2B, 4B, and 8B summary. A heterogeneous star with a CLIP image hub, a CLIP text spoke, and a Qwen image spoke is the same test when the encoder spaces do not already share a linear readout. Figure 4 reports spoke to spoke Recall@1 0.485 ± 0.131 against chance 0.125, with disconnected and shuffled controls at 0.212 and 0.172.

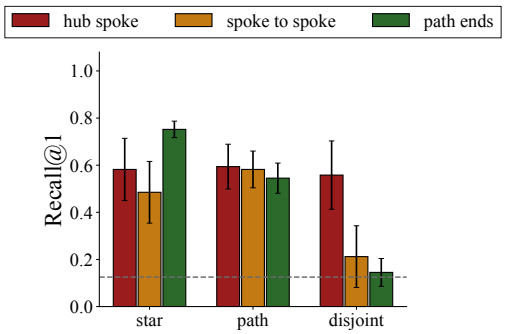


Figure 4: Heterogeneous star with a CLIP image hub, a CLIP text spoke, and a Qwen image spoke. Spoke to spoke Recall@1 is 0.485 ± 0.131 against chance 0.125. Disconnected and shuffled controls sit at 0.212 and 0.172.

Path end retrieval on CLIP is 0.624 ± 0.065 for the star and 0.455 ± 0.061 for the path. A nonlinear tanh world with known factors, trained by the same loss, hides residual z_{TA} on a star (0.061 and 0.134) and recovers it after the missing edge is added (0.629 and 0.630). Holding $M = 8$ and seven tree edges, worst pair MSE tracks $1/\lambda_2$ with $R^2 = 0.997$.

O.5 MCC VERSUS IDENTITY GAUGE

Hungarian mean correlation maximises $\frac{1}{d}\sum_i |\mathrm{Corr}(\hat{u}_i, z_{\pi(i)})|$ over permutations π. If two disconnected components each recover z_s up to independent diffeomorphisms, or up to independent signs, Hungarian matching will align those signs and report a large score, which hides the necessity half of Theorem 1. Identity gauge correlation compares coordinates in the order produced by glueing along a spanning tree, without a further permutation. Table 5 reports both scores. Two disjoint edges have Hungarian alignment 0.892 and identity gauge alignment 0.446, and the main text uses the latter. Panel (d) of Figure 3 overlays the two scores, and Figure 5 shows identity gauge alignment across graphs.

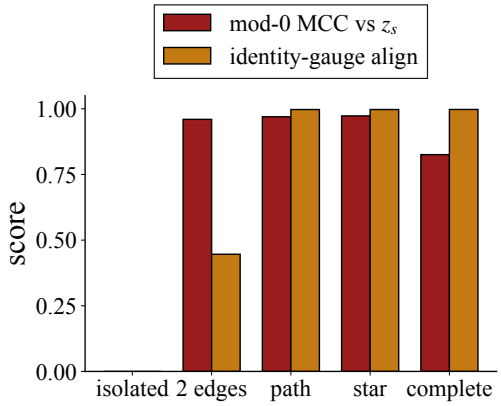


Figure 5: Identity gauge alignment across pairing graphs.

O.6 CONNECTIVITY OF THE PAIRING GRAPH

The linear Gaussian check uses four modalities and $d_s = 2$. Table 5 reports the mean correlation of each modality against z_s, together with the two consistency scores.

Table 5: Connectivity of the pairing graph. Identity gauge alignment is the consistency diagnostic.

graph	λ_2	MCC (mod 0)	Hungarian	identity gauge
isolated vertex	0	0.000	0.000	0.000
two disjoint edges	0	0.960	0.892	0.446
path	0.586	0.970	0.997	0.997
star	1	0.973	0.997	0.997
complete	4	0.826	0.997	0.997

An isolated vertex cannot separate shared factors from private factors given only a marginal, and its mean correlation is 0. Each connected component of the two edge graph recovers z_s internally, with mean correlation near 0.96, but cannot fix a relative gauge. The path, the star, and the complete graph align in the identity gauge at 0.997. Mean correlation on the complete graph is slightly lower, at 0.826, because the shared subspace is estimated from more noisy cross blocks, while consistency after glueing remains 0.997.

O.7 CLOSED-FORM λ_2

Table 3 compares the numerical Fiedler value of the unweighted Laplacian with Lemma 7. The numerical check is placed in Appendix N, next to the closed form identities.

O.8 HUB SUBSPACE INTERSECTION

Table 6 records MCC of the hub intersection against each latent block.

The one dimensional intersection of the two hub side CCA subspaces recovers z_s and is orthogonal to z_{TA}, with mean correlation 0.010. CCA on an additional text and audio edge recovers z_{TA} at mean correlation 0.912. The image and text edge recovers the larger block (z_s, z_{IT}) at the

Table 6: Hub subspace intersection, with and without a text and audio edge.

representation	z_s	z_{IT}	z_{IA}	z_{TA}
star intersection at I	0.999	0.001	0.004	0.010
T from edge IT	0.997	0.996	0.005	0.009
A from edge IA	0.867	0.006	0.869	0.021
star $+ TA$ edge	0.908	0.005	0.018	0.912

text encoder, which matches Corollary 1. A single spoke edge identifies $S_T \cap S_I$, not the global intersection.

O.9 Laplacian MSE on trees

Four trees on $M = 10$ vertices, each with nine edges and equal sample sizes, are compared in Table 7. Least squares mean squared error tracks $\mathrm{tr}(L^+)$ and $1/\lambda_2$ with Spearman correlation 1.00. The Gaussian formula $\sigma^2 \mathrm{tr}(L^+)/(nM)$ matches the empirical error to a few percent, namely 2% on the path, 4% on the star, less than 0.1% on the balanced tree, and 8% on the double star.

Table 7: Four trees on $M = 10$ vertices with nine edges. Empirical least squares error against $\mathrm{tr}(L^+)$.

tree	λ_2	$\mathrm{tr}(L^+)$	emp. MSE	theory MSE
path	0.0979	16.50	0.00404	0.00413
star	1.000	8.10	0.00212	0.00203
balanced	0.177	12.70	0.00317	0.00317
double star	0.298	9.70	0.00224	0.00242

O.10 Pairwise effective resistance

On a path with ten vertices, pairwise least squares error against effective resistance, which equals graph distance, has Spearman correlation 0.85. Finite sample fluctuations at distances 6 and 7 pull the rank correlation below the 0.98 observed on an earlier draw with nine vertices, while the slope remains positive. Resistance between two leaves of the star is exactly 2. Table 8 lists the pairwise error against R_{eff}, and Figure 6 plots the same pairs.

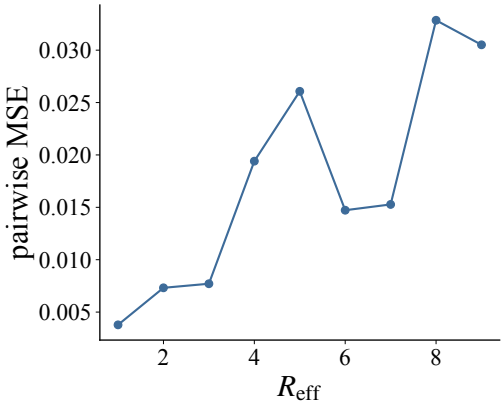


Figure 6: Pairwise MSE versus effective resistance on a path.

O.11 Hub versus path

The closed form for λ_2 agrees with the numerical spectrum to absolute error below 10^{-14} for every $M \in \{4, \dots, 12\}$ and for the star, the path, the cycle, and the complete graph. The worst pair least squares error grows with M on the path and stays flat on the star. The complete graph is lower still, because it has more edges, so the comparison is not at a fixed budget. At $M = 4$ the path and the star have similar worst pair error, since the path has only three hops. By $M = 12$ the worst pair error is 0.103 on the path and 0.051 on the star.

Table 8: Path pairwise MSE versus R_{eff}.

distance	R_{eff}	pairwise MSE
1	1	0.0038
2	2	0.0073
3	3	0.0077
4	4	0.0194
5	5	0.0261
6	6	0.0147
7	7	0.0153
8	8	0.0329
9	9	0.0305

O.12 Private factors, anchors, and leakage

Two view CCA recovers the shared block at mean correlation 0.980 and the private block at 0.016, with canonical correlations 0.998 and 0.997. Table 9 and Figure 7 report private mean correlation against the number of linear anchors.

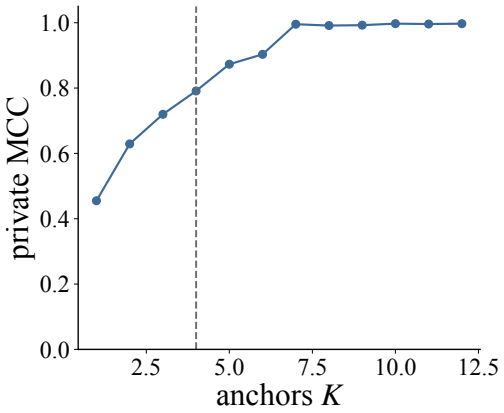


Figure 7: Private MCC versus the number of linear anchors.

Table 9: Linear anchors, $d_p = 4$.

K	private MCC	$K < d_p$?
1	0.455	yes
2	0.629	yes
3	0.720	yes
4	0.791	no
5	0.873	no
6	0.903	no
7	0.995	no
8 to 12	> 0.991	no

The counting lower bound $K \geq d_p$ is a population statement for linear anchors. Finite samples need a few extra anchors before the mean correlation exceeds 0.99, and Table 9 reports that curve rather than treating $K = 4$ as already saturated. Gaussian mutual information between a content encoder and the private block is $4.0 \cdot 10^{-4}$ nats, against 6.01 nats for the shared block. The raw observation has mutual information 8.99 nats with the private block and 6.01 nats with the shared block. The content encoder therefore leaks essentially nothing about the private block, while any invertible embedding of X_m leaks the full private entropy under the Gaussian formula.

O.13 Unpaired Brenier versus paired barycentre

Table 10 and Figure 8 compare unpaired Brenier with the paired barycentre on the diagonal Gaussian world.

The gap in private mean correlation, Brenier minus barycentre, is 0.483. Unpaired optimal transport recovers both axes and does not mark which axis is shared. The paired barycentre puts its mass

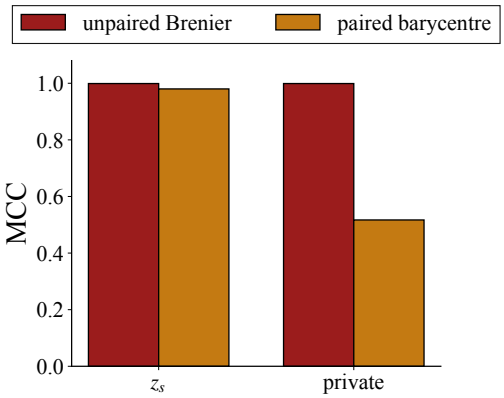


Figure 8: Unpaired Brenier map against the paired barycentre. The Brenier map mixes the private axis, and the paired map does not.

Table 10: Unpaired Brenier map and paired barycentre on the diagonal Gaussian world.

map	MCC vs z_s	MCC vs private
unpaired Brenier	1.000	1.000
paired barycentre	1.000	0.517
Brenier coord. 0 vs z_s	1.000	n.a.
Brenier coord. 1 vs private	n.a.	1.000
barycentre coord. 1 vs private	n.a.	0.144

on the shared axis, and the private mean correlation of its second coordinate is 0.144. The residual Hungarian private score of 0.52 is leftover correlation in a two dimensional map, not identification of z_p.

O.14 Hub SNR and the spiked-covariance threshold

Figure 9 reports a scan of hub signal strength on a star with strong spokes. The population correlation stays positive even at $\eta_h = 0.01$, and with large n the CCA direction remains recoverable, with hub mean correlation 0.887 at $\eta_h = 0.01$. That scan is not a collapse phase transition. It is included so that a plot with hub mean correlation 0.89 is not read as Theorem 5.

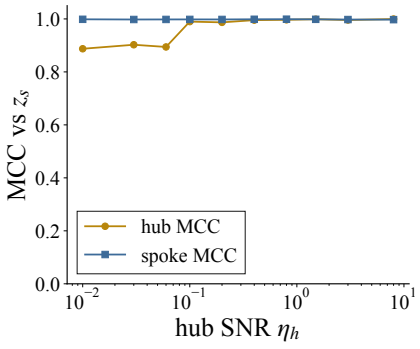


Figure 9: Hub SNR on a strong star. Hub MCC stays large whenever population $\rho > 0$.

The identity in Theorem 5 is an MMSE statement, checked in Tables 15 and 16. Operational collapse at finite sample size is the spiked scan in Table 11, which uses spiked CCA with $n = 300$, $D = 40$, and 25 trials. Figure 10 shows the same scan with error bars.

Below spike 1, the mean correlation stays between 0.04 and 0.08, and at spike 8 it rises to 0.57. The vertical line in Figure 10 is the order one spike threshold of Perry et al. (2018). It is not a fitted constant claimed as a theorem.

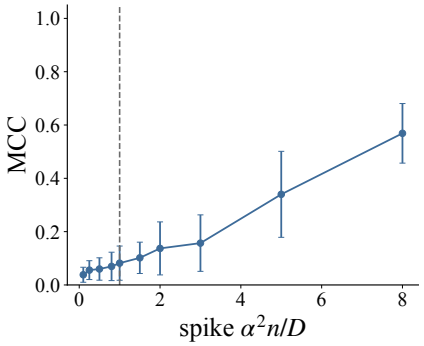


Figure 10: Spiked covariance mean correlation. The vertical line is the order one spike threshold.

Table 11: Mean MCC versus spike $\alpha^2 n/D$ (25 trials).

spike	mean MCC	std
0.10	0.038	0.028
0.25	0.056	0.035
0.50	0.060	0.042
0.80	0.070	0.053
1.00	0.082	0.064
1.50	0.102	0.059
2.00	0.137	0.099
3.00	0.157	0.106
5.00	0.340	0.161
8.00	0.569	0.112

O.15 Pairing-noise excess MSE

With $\eta_j = 2$ and a sweep in ε, the Spearman correlation of excess mean squared error against ε^2 is 1.00. The closed form $(1+\eta\varepsilon^2)/(1+\eta+\eta\varepsilon^2) - 1/(1+\eta)$ tracks the empirical curve in Figure 11 and in Table 12.

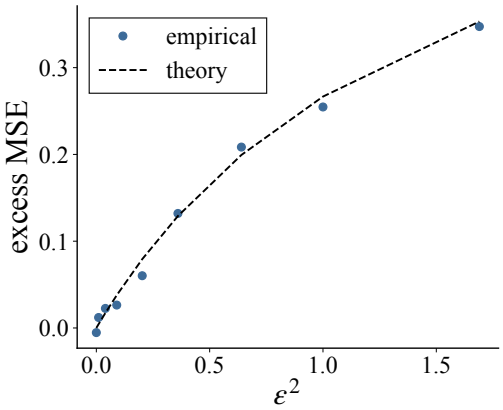


Figure 11: Excess MSE versus pairing noise.

O.16 Homophily without a pairing edge

There is no direct pairing edge between the two modalities, and neighbours on a data graph share z_s up to noise δ. Table 13 records canonical correlation against that homophily noise. Mean correlation is not the reported score, because it stays high at moderate δ. A strong shared direction and a large sample still find a correlated axis, which would overstate the theorem.

Table 12: Excess MSE versus pairing noise.

ε	ε^2	emp. excess	theory excess
0.0	0.00	-0.005	0.000
0.1	0.01	0.012	0.004
0.2	0.04	0.023	0.017
0.3	0.09	0.026	0.038
0.45	0.20	0.060	0.079
0.6	0.36	0.132	0.129
0.8	0.64	0.208	0.199
1.0	1.00	0.255	0.267
1.3	1.69	0.347	0.353

Table 13: Canonical correlation against homophily noise when pairing is induced by neighbours.

δ	MCC vs z_s	canonical correlation
0.00	0.999	0.997
0.25	0.999	0.965
0.50	0.999	0.897
1.00	0.999	0.705
2.00	0.998	0.435
4.00	0.997	0.247
independent	0.759	0.024

O.17 Two-view random-feature check

Random Fourier features followed by CCA, under tanh mixing, give paired mean correlation 0.288 and alignment 0.703. The shuffled pairing gives mean correlation 0.027 and alignment 0.386. The sign matches Theorem 1, and the quantitative nonlinear retrieval probe is Table 14.

O.18 Nonlinear class-conditional retrieval

The retrieval probe uses fixed tanh mixings, 10 classes, four modalities, and independent samples on each edge, with $n_{\text{edge}} = 2500$, $n_{\text{eval}} = 800$, and CCA rank $k = 8$. Chance Recall@1 is 0.1. Spoke to spoke retrieval on the star uses pairs that never co-occurred in training. Table 14 reports Recall@1 under the pairing graphs.

Table 14: Recall@1 under pairing graphs (chance 0.1).

graph	paired / adjacent	unpaired / distant	mean pair	connected
star	1.000	1.000 (spokes)	1.000	yes
path	1.000	0.889 (ends)	0.900	yes
two edges	1.000	0.035 (cross)	0.374	no
complete	1.000	0.999	1.000	yes

Two view controls on the same world give CCA Recall@1 of 1.000 when paired and 0.144 when shuffled. Alignment with one hidden layer of rectified linear units gives 0.999 when paired and 0.049 when shuffled.

O.19 Sample size on connected graphs

On the path, the star, and the cycle with six vertices, mean correlation is already above 0.99 at $n = 80$ per edge. The scan checks that connected graphs do not degrade as n grows. It is not a tight empirical plot of the rate $1/(n\lambda_2)$. That tightness is Table 7, where the estimator is exactly Laplacian least squares.

O.20 MMSE identities

Tables 15 and 16 check the single view and two view MMSE identities, and Figure 12 plots the single view curve against $1/(1 + \eta)$.

Mean relative error is 1.4% for one view and 1.7% for two views.

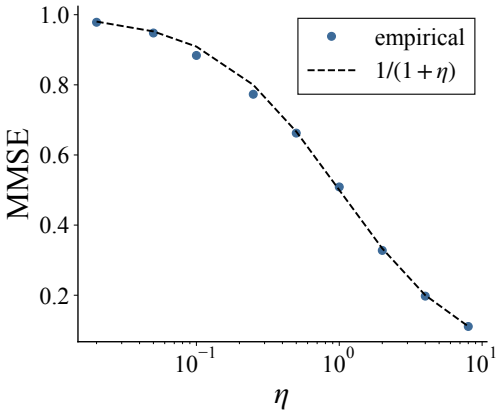


Figure 12: Closed form single view MMSE against $1/(1+\eta)$.

Table 15: Single view MMSE against $1/(1+\eta)$.

η	empirical MSE	$1/(1+\eta)$
0.02	0.978	0.980
0.05	0.948	0.952
0.10	0.884	0.909
0.25	0.773	0.800
0.50	0.662	0.667
1.00	0.509	0.500
2.00	0.328	0.333
4.00	0.197	0.200
8.00	0.111	0.111

Table 16: Two view MMSE against $1/(1+\eta_i+\eta_j)$.

η_i	η_j	empirical MSE	theory
0.1	0.1	0.829	0.833
0.1	2.0	0.309	0.323
0.5	0.5	0.493	0.500
1.0	1.0	0.341	0.333
0.05	4.0	0.200	0.198
2.0	2.0	0.198	0.200