SAME WORLD, DIFFERENT GEOMETRY: A CONTROLLED TEST OF THE PLATONIC REPRESENTATION HYPOTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Platonic Representation Hypothesis (Huh et al., 2024) holds that models trained on different modalities converge, with scale and competence, toward a common representation of the world. Testing whether models converge to the world, and not only to each other, requires knowing the world's true structure, which is unknown for natural images and text. Consequently, existing evidence compares representations only with each other, not with the world, and cannot manipulate what makes them agree. We therefore construct a world whose structure is known: eight Hamiltonian systems with ground-truth latent variables h, rendered independently as video and text. The two encoders share no weights or joint objective and are trained in matched pairs that differ only in their prediction targets, so the cause of alignment can be manipulated directly. We show that a shared world is not enough: each encoder recovers the physics, yet the two do not align. Alignment arises only when both are also trained to predict information about h. We characterize this phenomenon through the H-subspace, the subspace of each latent on which both encoders carry the same linear statistics of h. We prove that the latent spaces align on it and show that a single linear map transports information between the spaces. Alignment depends on the information the targets carry about h, not on whether they are shared: different targets, each informative about h, align the encoders, while an identical target uninformative about h does not. Large-scale pretraining does not substitute for this information: models pretrained on unrelated natural corpora decode the physics but share no such subspace. In this setting, convergence is *installed* by information about h, not by observing the same world.

1 INTRODUCTION

Do neural networks trained separately on different views of the world converge toward a shared representation of it? The Platonic Representation Hypothesis (PRH; Huh et al., 2024) answers yes: with enough scale and competence, models converge across architectures, tasks and modalities toward a common model of reality. The evidence so far is split. Convergence has been reported to grow with scale (Huh et al., 2024), but also to depend on the regime (Koepke et al., 2026) and on the metric (Gröger et al., 2026), and to saturate well below ceiling (Zhu et al., 2026). The disagreement is hard to resolve because the measures in use (CKA, mutual-kNN, retrieval) compare representations only with one another. For natural images and text the true generative structure is largely unknown, so one cannot check what the models are converging *to*, or intervene on the cause.

We give the PRH what is, to our knowledge, its first controlled test, in a world whose structure is known: eight classical-mechanical systems with a ground-truth latent h. Each trajectory is rendered as a video and described in a natural-language caption, and a separate generative model is trained on each modality. The two encoders are trained in complete isolation, so any convergence must come from the data and the per-modality objectives. Because h is known, we can also manipulate what the PRH leaves implicit: what each model is asked to encode. The only prior controlled study of cross-modal alignment couples the encoders through a shared contrastive objective (Gupta et al., 2026); we remove that coupling.

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.

Table 1: The experiments, proofs and analyses behind each contribution, and where they appear.

Contribution	Main text	Appendix
The testbed: eight Hamiltonian systems with known h, rendered as video and captions along a four-level disclosure ladder, two held-out systems; the alignment metrics and their invariance classes.	§3, §5	App. A, B, C
Matched pairs with and without the reference (eight systems, five seeds for the main pairs); shuffled, disjoint, asymmetric and partial-overlap targets; coupling baselines; frozen pretrained pairs.	§6.1, §6.3, §6.4	App. D, F, G.6
The H-subspace definition, the two training problems, and the theorem identifying the subspace with the fixed space of the cross-modal round trip; its measured dimension against the shared target.	§4, §6.3	App. I, F
Cross-modal maps (linear, nonlinear and invertible), round trip, parameter transfer and stitching; sweeps over scale, data, read-out and objective; zero-shot evaluation on three-body and Coulomb systems.	§6.2, §6.4	App. E, G, H

Trained on the same world, each encoder recovers the physics, yet the two latent spaces do not align. They align only when both are also trained to predict information about h, which then acts as a common reference. A double dissociation shows that shared information, rather than a shared task, drives this: an identical target uninformative about h leaves the spaces apart, while two different targets, each informative about h, align them substantially.

We characterize this through the H-subspace, the part of each latent on which the two encoders carry the same linear statistics of h, and prove that the spaces align there. Without a reference this subspace is nearly empty: the encoders agree coarsely but rarely identify individual samples. With one, it grows with the information the two targets share. Since the PRH attributes convergence to scale, we also test whether large-scale pretraining can substitute for the reference. Two models pretrained on unrelated natural corpora decode our physics from frozen features but share no H-subspace, while our encoders trained without a reference still share a small one.

This suggests an explanation for the regime dependence reported in prior work: models trained on the same world can still encode it in unrelated frames, so alignment should appear where training pipelines tie modalities to a common reference, such as labels, captions or distillation, and fade where they do not.

Our contributions.

- We present, to our knowledge, the first controlled test of the PRH against a known ground-truth latent, via a reusable testbed: eight Hamiltonian systems rendered independently as video and text, with held-out systems.
- We show that alignment depends on information about h: it grows with the information the two targets share, while models that learn h without such a target, including large pretrained ones, do not align.
- We characterize the resulting correspondence through the H-subspace, the subspace on which both encoders carry the same linear statistics of h, and prove that a single linear map carries one onto the other there.
- We show that the induced code transfers near-losslessly: one linear map returns each point to itself, carries what one modality encodes about a sample into the other, and transfers zero-shot to a held-out force law.

2 Related Work

2.1 Shared World

Cross-modal alignment has long been built explicitly into training objectives, from contrastive predictive coding (van den Oord et al., 2018) to contrastive vision–language models such as ConVIRT (Zhang et al., 2022), CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). The hypothesis was then generalized by the PRH (Huh et al., 2024), which argues that architectures silently con-

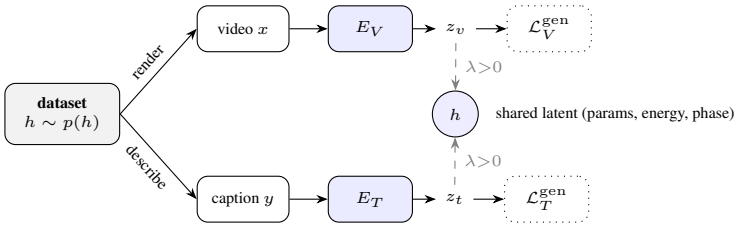

Figure 1: **The shared-reference test.** A latent h generates both video x and caption y; the encoders never interact; at $\lambda>0$ both additionally regress to h.

verge across modalities, tasks, and data. Subsequent work studies this convergence through training dynamics (Ziyin et al., 2025), data (Tjandrasuwita et al., 2025) and optimizers (Ziyin & Chuang, 2025), while others challenge its metrics (Gröger et al., 2026) or argue that it is regime-dependent (Koepke et al., 2026). Closest to our thesis, Vafidis et al. (2025) show that multi-task training recovers block-identifiable, task-relevant subspaces. Closest in construction, Gupta et al. (2026) relate two independently trained CLIP-style models by a single orthogonal map and prove that agreement of the multimodal kernel on a small anchor set forces an isometry. Their alignment, however, comes from a shared cross-modal contrastive objective. Our encoders are trained with no cross-modal term at all, which isolates whether a common target on h can induce alignment by itself.

A parallel line aligns independently trained encoders *post hoc*: their geometries often differ by little more than a linear map (Maniparambil et al., 2024; Merullo et al., 2023), yet recovering that map without paired data is hard (Schnaus et al., 2025; Jha et al., 2025) and succeeds mainly with paired anchors (Moschella et al., 2023; Norelli et al., 2023; Hadgi et al., 2025). Deep CCA and multi-view objectives such as VICReg and Barlow Twins (Andrew et al., 2013; Bardes et al., 2022; Zbontar et al., 2021), induce shared factors across views without labels, and identifiability theory shows that such joint objectives recover them up to an invertible transformation (Gresele et al., 2019; von Kügelgen et al., 2021; Daunhawer et al., 2023). These guarantees rely on a loss that couples the encoders, which our setting excludes. Most similar in setting, Zhu et al. (2026) find video–text alignment between independent encoders depends on test-time data richness, reaching about 0.4 mutual-kNN. Together these works locate where same-data alignment can fail: within a modality, independently trained models are often related by a linear map (Roeder et al., 2021; Bansal et al., 2021), whereas across modalities, with only the data shared, we find that no map recovers a trajectory-level correspondence.

2.2 Learned Dynamics

Physical systems are a natural testbed: every observation is generated by a few known quantities, parameters and initial conditions, under fixed laws. Several works build the physical structure into the network: Hamiltonian and Lagrangian Neural Networks (Greydanus et al., 2019; Cranmer et al., 2020) and Symplectic ODE-Net (Zhong et al., 2020) make conservation structural, Hamiltonian Generative Networks (Toth et al., 2020) learn such dynamics from pixels, and SyMetric (Higgins et al., 2021) measures how faithfully learned dynamics recover the true phase space. We instead build no physics into the encoders, using the known Hamiltonian only as a ground-truth target.

3 Problem Formulation

We work in the framework of data pairs $\{(x, h), (y, h)\}$, where x is a video, y is text, and h is the full latent that generates both: the Hamiltonian parameters θ and the initial condition (q_0, p_0), which together determine the trajectory and its energy, and the appearance (color, object count). The data is generated by a common cause: a truth $h \sim p_{\mathcal{H}}$ is drawn, and the video and caption are rendered from the resulting trajectory independently,

$$p(x, y, h) = p(h)\, p(x \mid h)\, p(y \mid h), \qquad \text{i.e.} \qquad x \perp y \mid h, \tag{1}$$

so in the fork $x \leftarrow h \rightarrow y$ all statistical dependence between the two modalities flows through h.

We study the encoders of two separately trained models, $f : \mathcal{X} \to E \subseteq \mathbb{R}^d$ (video) and $g : \mathcal{Y} \to E' \subseteq \mathbb{R}^d$ (text), where $f(x), g(y)$ are the pooled read-outs of the video backbone and the autoregressive text model. Each model optimizes a purely generative objective on its own modality,

$$\mathcal{L}^V_{\text{gen}}(f) = \mathbb{E}\big[\|\epsilon_f(x_\tau, \tau) - \epsilon\|_2^2\big], \qquad \mathcal{L}^T_{\text{gen}}(g) = \mathbb{E}_y\Big[-\textstyle\sum_t \log p_g(y_t \mid y_{<t})\Big], \tag{2}$$

where the first expectation is over x, τ uniform over the diffusion steps, and $\epsilon \sim \mathcal{N}(0, I)$, with $x_\tau = \alpha_\tau x + \sigma_\tau \epsilon$: a denoising objective for the video model and an autoregressive objective for the text model. Neither references h nor the other modality.

In addition we define an auxiliary objective that ties each latent to the shared truth through light physics heads $\rho_V : E \to \mathcal{H}$, $\rho_T : E' \to \mathcal{H}$,

$$\mathcal{L}^V_{\text{aux}} = \mathbb{E}_{(x,h)}\big[\ell\big(\rho_V(f(x)),\, h\big)\big], \qquad \mathcal{L}^T_{\text{aux}} = \mathbb{E}_{(y,h)}\big[\ell\big(\rho_T(g(y)),\, h\big)\big], \tag{3}$$

where ℓ is a masked regression loss onto h. Each model is trained with

$$\mathcal{L}^V = (1-\lambda)\mathcal{L}^V_{\text{gen}} + \lambda\, \mathcal{L}^V_{\text{aux}} \tag{4}$$

so $\lambda{=}0$ recovers the purely generative models and $\lambda{>}0$ imposes the reference h (schematic; the implementation is a binary gate with an adaptive balancer, App. B). In the default configuration the two heads regress the same h; §6.3 varies *what* each head is asked to predict, decoupling the information a target carries from whether the two encoders receive the same target.

We now turn to the latent spaces E, E', both of dimension d, with pushforwards $p_E = f_\# p_{\mathcal{X}}$ and $p_{E'} = g_\# p_{\mathcal{Y}}$ (the laws of $e = f(x)$ and $e' = g(y)$). The conditional independence of Eq. (1) carries to the latents, $e \perp e' \mid h$, and marginalizing the cause yields the coupling on $E \times E'$ induced by h, $\pi(e, e') = \int_{\mathcal{H}} p(e \mid h)\, p(e' \mid h)\, dp_{\mathcal{H}}(h)$, whose marginals are $p_E, p_{E'}$. We formalize our problem as the alignment of the two spaces through this coupling: they are aligned if there exist a modality-agnostic form $\mu : \mathcal{H} \to \mathbb{R}^d$ and a transform T in a class $\mathcal{T}$ (specified below) such that the per-cause conditionals agree,

$$p(e \mid h) \approx T_\# \, p(e' \mid h) \quad \forall h, \qquad \text{with} \qquad \mathbb{E}[e \mid h] \approx \mu(h), \quad \mathbb{E}[e' \mid h] \approx T^{-1}\mu(h), \tag{5}$$

i.e. both encoders express the same latent geometry over $\mathcal{H}$ up to T; we take $\mathcal{T}$ to be the general invertible linear maps our fitted maps realize (§6.2; how each metric's invariance class relates to $\mathcal{T}$ in App. C). Our central question is whether Eq. (5) holds from the generative objectives alone ($\lambda{=}0$) or only under the reference ($\lambda{>}0$): whether alignment is emergent or induced, and, if induced, whether by the information the targets carry about h or by the sharing of an objective (§4 formalizes both).

4 The H-Subspace: A Characterization of Alignment

Let $e = f(x) \in E$ and $e' = g(y) \in E'$ denote the two latents, jointly distributed under the coupling induced by h, with $\dim E = \dim E' = d$.

Definition 1 (H-subspace)**.** Assume nonsingular latent covariances. The *H-subspace* is the pair of subspaces $\mathcal{S}_V \subseteq E$, $\mathcal{S}_T \subseteq E'$, of common dimension $k \le d$, on which the two latents carry the same content: there exist linear read-outs $U, V : \mathbb{R}^d \to \mathbb{R}^k$ of rank k with $Ue = Ve' = c(h)$ almost surely for some statistic $c : \mathcal{H} \to \mathbb{R}^k$, taken maximal (collecting every such statistic), and $\mathcal{S}_V = \operatorname{col}\big(\operatorname{Cov}(e, c)\big) \subseteq E$, $\mathcal{S}_T = \operatorname{col}\big(\operatorname{Cov}(e', c)\big) \subseteq E'$ are the subspaces in which $c(h)$ is embedded, with U, V their read-outs. On the H-subspace the two latents are *globally aligned*: a single linear bijection $\mathcal{S}_V \to \mathcal{S}_T$ carries one onto the other (Theorem 7(iii)).

The definition places no restriction on the form of c: it is whatever both latents happen to encode linearly about h, and k may be zero. Our claim is that at $\lambda{=}0$ the H-subspace is nearly empty, and the reference fills it. The two terms of Eq. (4) act on different parts of the latent (Figure 2).

Definition 2 (The two training problems)**.** Each modality has a *target*, a statistic $s : \mathcal{H} \to \mathbb{R}^{d_h}$ of h, $d_h \le d$, that its head ρ_V or ρ_T predicts: the full h, a single quantity such as the energy, or different quantities for the two modalities (§6.3). The *generative* problem trains each encoder with a decoder on reconstruction alone and engages all d dimensions,

$$(P^0): \qquad \min_{(f, D_V)} \ \mathbb{E}\left\|x - D_V(f(x))\right\|_F^2, \qquad \min_{(g, D_T)} \ \mathbb{E}\left\|y - D_T(g(y))\right\|_F^2. \tag{6}$$

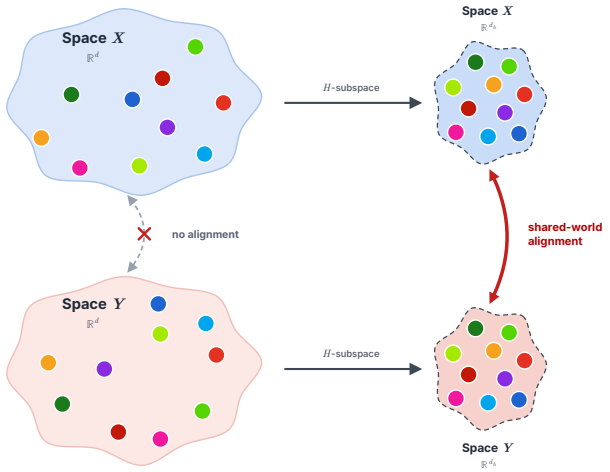

Figure 2: **The mechanism.** Left: trained alone (λ=0), no alignment. Right: the reference installs the information subspaces on h (dashed), where alignment lives (Theorem 7).

The *reference* problem concerns d_h of them: f is H-*sufficient* if $E = \mathcal{S} \oplus \mathcal{N}$ with $\dim \mathcal{S} = d_h$ and the $\mathcal{S}$-component of $f(x)$ is an affine bijection of $s^*(x) := \mathbb{E}[\, s(h) \mid x \,]$, the information the observation carries about the target. Its distance from this property, over affine probes $\pi : \mathbb{R}^d \to \mathbb{R}^{d_h}$, is

$$\mathrm{dec}_H(f) \;:=\; \inf_{\pi} \mathbb{E}\big\|\pi(f(x)) - s(h)\big\|_2^2, \qquad (P^H): \;\; \min_{f} \; \mathrm{dec}_H(f), \tag{7}$$

whose solutions, for $\mathrm{Cov}(s^*) \succ 0$, are exactly the H-sufficient encoders (App. I.1).

With (P^0) and (P^H) as its generative and auxiliary terms, Eq. (4) becomes their weighted combination, a *double requirement*; for $\lambda \in (0, 1)$ the minimizers of this combination are Pareto optimal, no encoder improving one risk without worsening the other (Boyd & Vandenberghe, 2004). A trained encoder therefore places the target's information in its subspace as far as reconstruction permits.

Trained on (P^0) alone, each encoder lays out its latent as it likes, spreading what it knows about h across whatever dimensions it chooses, and two such layouts share only what both modalities happen to display. The reference does not change what the encoders may do with E and E'; it places what each observation says about h in a d_h-dimensional subspace of its latent, not in the head that reads it (App. B, F.2; Table 11), and since the target includes the per-frame phase, the subspaces determine the trajectory up to appearance (Figs. 3, 7). These subspaces are the H-subspace, and on them the latents are aligned: a single linear map carries one onto the other, and the composition of the two fitted cross-modal regressions returns these directions unchanged (Theorem 7, App. I). The alignment measured on the two latent spaces is carried by this subspace: retrieval, round trip and transfer follow from the content it holds, while directions outside it are matched only partially ($\rho^2 < 1$) and the map is not the identity there (App. I.2). The aligned part grows with what the targets share: beyond the small floor present at λ=0, targets with no common component add no directions and raise alignment only through such partial correlations (App. I.4), a scalar anchor adds at most one, and the full target adds as many as its effective dimension (Table 32). This does not depend on the generative objective: the same dissociation holds across objective families (App. G.4). Appendix I gives the proofs, a concentration bound for independently chosen frames, the empirical location of the subspace, and the statement of Eq. (24).

5 Experimental Setup

5.1 Dynamical Systems

Our simulation and rendering build on DeepMind's Hamiltonian Dynamics Suite (Botev et al., 2021), which provides the base systems, the integrator, the renderer, and the within-canvas rejection

sampling; our additions are a three-body problem, a charged (Coulomb) two-body system with a second, sign-varying force law, and the natural-language caption ladder. The systems span mass–spring, pendulum, double pendulum, and two- and three-body gravity plus the charged variant. Tables 3–4 (App. A) give each system's equation, parameters, regime, and varied quantities. The point-mass systems are separable Hamiltonians $H(q,p) = \sum_i \|p_i\|^2/2m_i + U(q;\theta)$ with the force law in the potential; the pendula use the generalized-coordinate Hamiltonians of Table 3. The four base systems and three-body add *damped* variants ($\dot{p} = -\partial_q H - \eta\,\dot{q}$, monotonically dissipating H; η=0 recovers the conservative system), and h records whether a trajectory is damped.

5.2 Modalities and Encoders

Each h determines a trajectory, which we integrate and subsample to 61 frames at Δt=0.05, with initial conditions rejection-sampled to stay in-frame, and render as a video and a text description, giving the pair $\{(x,h),(y,h)\}$. The video draws the bodies as colored discs (radius encoding mass) on a fixed 32×32 frame; per system we render 7500 trajectories (6000 train, 1500 test). The text describes the same trajectory along a four-level disclosure ladder L0–L3 of raw observations and visible attributes, one deterministic templated caption per level (240,000 captions, 192,000 train); the energy is withheld at every level and must be inferred from the observations. Solver, rendering, per-level contents and example captions are in App. A.

The video encoder is a Latte diffusion transformer (Latte-S/4; Ma et al., 2025), built on DDPM (Ho et al., 2020), Latent Diffusion (Rombach et al., 2022) and DiT (Peebles & Xie, 2023), consuming 16 evenly subsampled frames per clip; the text encoder is a GPT-2 transformer (Radford et al., 2019) consuming captions up to 256 tokens, width-matched so both latent spaces have dimension $d = 384$. The read-outs $f(x), g(y)$ are either a mean pool of the hidden states or a learned K-query bottleneck that cross-attends to the sequence, giving a shared $\mathbb{R}^{Kd}$ latent on which generation is also conditioned (App. B). The heads ρ_V, ρ_T of Eq. (3) regress a fixed-width union of the parameters, the energy and the per-frame phase $(q,p)(t)$, masked per system to the active entries; appearance is part of h but never a target. Training details are in App. B.

5.3 Training and Evaluation

Every λ comparison in the paper is a matched pair: two encoders per modality trained on the same data with the same architecture and optimization, differing only in λ. Training runs 30 epochs on the eight base systems with the best validation checkpoint retained; the main pair and its target variants run five seeds each (App. B). Three-body and Coulomb are held out for out-of-distribution evaluation. The target variants, scale sweep and training-breadth sweep are specified where they are used (§6.3, §6.4).

All metrics are computed on frozen encoders over paired test trajectories, and **higher is better on all metrics**. Map-free alignment metrics compare the raw latent spaces: linear CKA (Kornblith et al., 2019), RSA (Kriegeskorte et al., 2008), SVCCA (Raghu et al., 2017), and mutual-kNN, the convergence metric of Huh et al. (2024). Fitted-map metrics use ridge maps $W_{v\to t}, W_{t\to v}$ fit on a $70/30$ split and evaluated held out: cross-modal retrieval (r@1, reported after the map; raw-space retrieval is near chance in both conditions), Procrustes (Schönemann, 1966), *transfer*, which reads a probe fit in one modality out of the other through the map, and generation, which stitches each encoder to the other modality's decoder (Fig. 4). *Decodability* is a post hoc ridge probe (Alain & Bengio, 2017) within one modality for each sub-block of h, distinct from the heads that shape the encoder. Permutation tests confirm no inflation by scale, dimensionality or sampling (App. D.2). Definitions and each metric's invariance class (Table 8) are in App. C.

6 Results

6.1 Cross-Modal Alignment Is Induced by the Shared Reference

The dissociation. Across all eight systems and five seeds, cross-modal CKA is 0.21 without the auxiliary task and 0.72 with it, a $3.4\times$ gap, and the identity-level measures separate far more sharply: mutual kNN, the convergence metric of the PRH (Huh et al., 2024), rises from 0.047 to 0.41 ($\sim 7\times$ and $\sim 60\times$ chance), and mapped r@1 from 0.042 to 0.84 (chance $\approx$ 0.002; Table 2). Without the

Table 2: Cross-modal alignment by target, mean±std over eight systems, each averaged over five seeds; higher is better, best in bold. Asymmetric targets: vision regresses energy, text parameters. The first two rows are frozen pairs of public checkpoints of matched architecture, pretrained on natural data, never on our systems; having no seeds, they give one value per system, and norm. CKA, whose ceiling needs seeds, is undefined. All four pretrained pairs are in Tables 43 and 44 (App. G.6).

target	CKA ↑	norm. CKA ↑	mutual-kNN ↑	mapped r@1 ↑
pretrained, Latte-B / GPT-2 small (d=768)	0.06±0.02	n/a	0.018±0.004	0.12±0.02
pretrained, Latte-L / GPT-2 medium (d=1024)	0.06±0.02	n/a	0.017±0.003	0.11±0.02
none (λ=0)	0.21±0.07	0.23±0.08	0.047±0.019	0.04±0.02
identical, independent of h	0.19±0.07	0.27±0.10	0.044±0.020	0.04±0.02
asymmetric, each informative about h	0.41±0.10	0.52±0.13	0.146±0.061	0.28±0.14
identical, all of h	**0.72±0.09**	**0.77±0.10**	**0.408±0.076**	**0.84±0.09**

target these fall close to chance while CKA keeps partial global structure, so the reference adds point-level correspondence on top of coarse similarity (Table 12). The distributions do not overlap, the worst auxiliary run beating the best control on every measure across all runs (Table 10), and both conditions reproduce across seeds (Table 9), so the floor is not seed noise.

Decodability does not explain alignment. Without the target both encoders already decode the parameters (ridge probe R^2 0.79 video, 0.85 text, against 0.96 and 0.95 with it; Table 11), and a linear map carries them across at $R^2 \approx 0.7$ (Table 18). Yet the spaces stay at the floor above, and retrieval stays near it even within the subspace spanned by the parameter probes (Table 15). Both latents therefore carry h, but in unrelated frames, a case the PRH does not consider.

6.2 Cross-Modal Maps and Their Structure

Four types of maps. We fit four maps (least-squares, linear layer, MLP and lightweight transformer) on training pairs, scored on *held-out* pairs (Table 16). With the auxiliary head a linear least-squares map already suffices (text→vision $R^2 = 0.77$, cosine 0.87, mapped r@1 0.84), with the MLP and transformer adding little; without the reference, no map class, including a broad nonlinear search (RBF ridge, kernel CCA, deep MLPs), recovers a trajectory-level correspondence (Table 20), so the failure lies in the representations.

The fitted maps are not rotations. The text→vision map deviates from orthogonality by 6.70 ($\|W^\top W/\|W\|_2^2 - I\|_F$), and whitening both spaces brings this down only to 1.87 (Table 25). This is what Corollary 6 predicts under Eq. (24): each encoder embeds $s(h)$ in its own frame, so the map between them is $R_V R_T^+$, a general linear map (Hyvärinen & Morioka, 2016; Khemakhem et al., 2020; Roeder et al., 2021; Reizinger et al., 2025), and an isometry only on the whitened H-subspace (Fig. 8). Prior work finds rotations because contrastive objectives whiten the representations (Gupta et al., 2026; Zimmermann et al., 2021; Torralba & Weiss, 2026), which our unnormalized objectives do not.

Round-trip, transfer, and symmetry. With the reference, the round trip returns each point to itself (cycle recall-at-1 1.00 text / 0.98 vision), although the matrix product is far from the identity ($\|WW - I\|/\|I\| \approx 1.65$; Table 17): as Proposition 5(ii) predicts under Eq. 24, the cycle is the identity only on the H-subspace. Transfer through the map is near-lossless (R^2 0.95; Table 18), and the aligned code follows the physics and ignores appearance: it is invariant to a color swap and equivariant to a mass swap (Figs. 3, 7). Without the reference, the cycles fall to 0.83 (text) / 0.43 (vision) and transfer to 0.75/0.68 (full analyses in App. E).

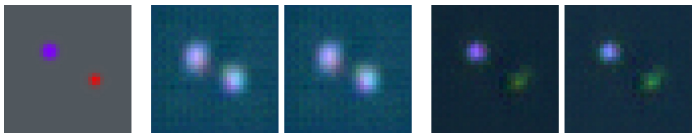

Figure 3: **The mass-swap counterfactual.** A caption with the masses swapped (big↔small) fed into D_V: ground truth (1); no-aux (2–3) and aux (4–5) as mapped latent and round trip. Only under the reference does the decoded scene follow the swap.

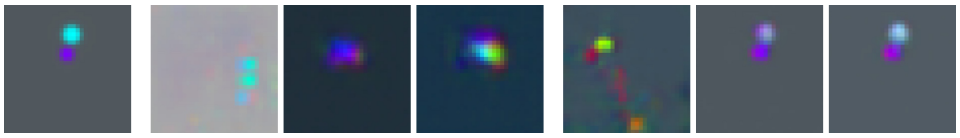

Figure 4: **Cross-modal stitching.** Ground truth (1); no-aux (2–4) and aux (5–7) decoded by the frozen D_V as direct text latent $(2, 5)$, mapped text $D_V(W_{t\to v} z_t)$ $(3, 6)$, and round-trip $(4, 7)$. A faithful scene appears only under the shared reference.

6.3 It Is Shared *Information*, Not a Shared *Objective*, That Aligns

Varying what the two encoders are asked to predict. We vary the target from no shared content to an identical one (Table 2; full results in Table 28). The variants form a double dissociation between what a target carries about h and whether the two encoders share it. A *shuffled target*[1], shared by both encoders but independent of h, leaves alignment at $\lambda{=}0$ level (CKA 0.19, r@1 0.04), while two different targets, each informative about h, raise it well above that level (CKA 0.41, r@1 0.28), short of the identical full target (0.72, 0.84).

Coupling baselines. Two non-contrastive multi-view objectives we test, VICReg and Barlow Twins, match the two latents directly with *no* ground-truth h, coupling the encoders as a CLIP-style loss would; both raise global similarity (CKA 0.54 and 0.68). They differ at the point level: VICReg's per-sample invariance term produces correspondence (r@1 0.35), while Barlow Twins, which matches only feature cross-correlations, does not (r@1 0.05; App. B), the same coarse-versus-identity split we draw between CKA and retrieval. The shuffled target, by contrast, is shared but does not couple the encoders.

A shared objective is not necessary. Conversely, the *asymmetric* pair aligns without receiving the same target: vision predicts only the energy and text only the parameters, two informative functions of h with no shared component (Table 28). Reversing the assignment aligns comparably (CKA 0.45, r@1 0.39; Table 30), so the effect does not depend on which modality receives which target.

Alignment tracks information. Between these poles, alignment grows with how much the target says about h. An information-preserving random *projection* of h aligns (CKA 0.48), whereas the shuffled target, shared to exactly the same degree, does not (Table 28). The strength of alignment follows the content of the target: a single scalar, the energy, lifts it clearly above the floor, while a categorical system label, constant within each evaluation set and therefore uninformative about h there, leaves it there (Table 29). A partial-overlap sweep then rises monotonically in CKA to the identical target (Table 30), which is the point of maximal shared information. Counting near-unit canonical correlations is consistent with this (Table 32; Theorem 7(ii)): the count sits at a small common floor whenever the targets share nothing, and rises as they share more information.

6.4 Robustness, Generalization, and Scale

The dissociation is robust along four axes. *Read-out:* bottleneck width and pooling method (Tables 35, 36); a compressed K-query pool raises both conditions, and the auxiliary margin persists at every width (Fig. 10). *Auxiliary objective:* removing the per-frame phase head, changing the loss

[1]This control shuffles the *correspondence* between target and trajectory, destroying information; a coordinate permutation of h would be invertible and information-preserving.

balance, reducing the target to a minimal or partial one (Tables 26, 31, 29, 30), and collapsing the head to a single affine map (CKA 0.74 vs. 0.69; App. F.2). *Data and evaluation:* caption length, neighbourhood size, and post-hoc Procrustes/CCA/ridge anchor fitting (Tables 37, 23, 22). *Models:* swapping the vision backbone (VideoFlow-S (Kumar et al., 2020); Table 38) or the text objective (autoregressive, score-entropy (Lou et al., 2024) or masked diffusion (Sahoo et al., 2024); Table 39), where matching the vision model's diffusion form lowers the $\lambda{=}0$ level rather than raising it. The reference does not improve generation (Table 27). Full analyses in Apps. E–H.

Zero-shot evaluation on the two held-out systems. On the chaotic *three-body* system the two conditions are indistinguishable in CKA (0.59 vs. 0.56; Table 12) but remain separated in retrieval (0.26 vs. 0.05), a separation that narrows gradually with chaos (Table 48). Generation itself is intact (Fig. 11). On the *Coulomb* system, which holds out the force law and not merely its parameters, the dissociation persists on every measure: with the reference, CKA is 0.41 and r@1 0.49 (from 0.72 and 0.84 in distribution), against 0.22 and 0.02 without it. The *asymmetric* target transfers partially (r@1 0.08/0.09; Table 46), the low-rank correspondence a scalar anchor predicts (App. I.4). Full details in App. H.

The gap is not an artifact of capacity or data. Scaling both encoders from Latte-S to Latte-B ($d{=}384/768$; Table 40) leaves the control at the floor (CKA $0.21 \rightarrow 0.23$) while the auxiliary model holds ($0.72 \rightarrow 0.72$), and Latte-L repeats the pattern (0.70 vs. 0.08). Neither 80k trajectories at fixed capacity nor two-to-eight-system training breadth moves the floor (Tables 41, 42; App. G.5).

Zero-shot pretrained pairs. Large-scale pretraining does not supply the reference either. We freeze four pairs of public checkpoints, Latte-B/L (Ma et al., 2025) trained on FaceForensics (Rössler et al., 2018), each paired with a GPT-2 (Radford et al., 2019) or SEDD (Lou et al., 2024) model of matched width, depth and heads; within a pair the two share no data, objective or tokenizer, and neither has seen our systems. Their frozen latents decode the Hamiltonian parameters ($R^2 \approx 0.85$ video, 0.53–0.65 text), yet every map-free measure falls below the $\lambda{=}0$ level (CKA 0.06–0.12 vs. 0.21) and no pair has a single canonical correlation at $\rho^2{>}0.9$, against 1.9 at $\lambda{=}0$ and 5.6 under the full target (Tables 43, 32). Retrieval alone rises (0.10–0.12), as a map fit on paired examples exploits their wider latents, but stays seven times below the reference. Our $\lambda{=}0$ encoders, trained on a common world, keep a small H-subspace, while the pretrained pairs have none (App. G.6).

7 Conclusion and Limitations

Conclusions. We built a controlled testbed for the Platonic Representation Hypothesis, in which the cause of cross-modal alignment can be manipulated rather than inferred. Across our experiments, two encoders trained on the same world recovered the physics without sharing a geometry, and they aligned once both were trained to expose information about h, with the aligned subspace growing with the information the two targets share. For the PRH, this suggests that convergence observed in the wild may reflect common references in training pipelines (labels, captions, distillation) as much as models independently arriving at a shared reality. It also points to a constructive route, in which an explicit reference for natural data, an agreed statistic of the world that independent models are trained to expose, would produce alignment across models and modalities up to the dimension of the information it carries.

Limitations. Our work has three main limitations. First, we do not include a CLIP-style contrastive baseline: its alignment is produced by a shared cross-modal objective and is already characterized (Gupta et al., 2026). Second, we measure alignment only on our synthetic systems; the pretrained pairs saw natural corpora in training, but not in evaluation. Third, our encoders are not frontier scale (12–457M parameters), whereas the PRH predicts convergence in the limit of scale; within our range, no data or model sweep narrows the gap (§6.4).

AI USE STATEMENT

In this work, we used generative AI tools for the following: Generating the templates for the matching captions, helping refine certain mathematical claims, provide feedback to experimental data and results, and refine text writing. We have not used generative AI tools for: the ideas, the direction, paper structure and hypothesis. Additionally we used AI tools for some of the code tasks, which were reviewed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work raises no direct ethical concerns. It involves no human subjects and no personal, private, or scraped data: every dataset is procedurally generated from known Hamiltonian systems (Botev et al., 2021), with videos and captions rendered deterministically from simulated trajectories (App. A). The study is methodological, asking what causes two independently trained encoders to converge. Its principal consequence is interpretive: representational convergence measured by similarity scores can reflect a shared supervisory reference in the training pipeline rather than models independently arriving at a common model of the world. We view this as a cautionary result for how convergence claims are read, and as potentially useful for auditing representational interoperability; we are aware of no dual-use or misuse risk specific to it. We report the compute used (App. B) in the interest of transparency about resource cost. Our work adheres to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have aimed to make every result reproducible. The data-generating process—the eight base systems, the two held-out systems, the integrator, the renderer, and the four-level natural-language caption ladder—is specified in full in App. A (Tables 3–4) and builds on the publicly available DeepMind Hamiltonian Dynamics Suite (Botev et al., 2021). Encoder architectures and widths (Latte-S/B/L and the width-matched GPT-2 text encoder), the auxiliary heads, the shared-reference target and its per-system mask, and all optimization settings are given in App. B (Tables 5, 6); the read-out variants in App. B. Every metric is defined in §5.3 and App. C, including the cross-modal map-fitting protocol (the $70/30$ train/test split). Unless noted, all reported numbers are mean$\pm$std over the five seeds $\{7, 42, 123, 1337, 2024\}$, and permutation tests for the similarity scores are in App. D.2. The code, data-generation scripts and experiment scripts are available in the supplementary materials. The links to download all pretrained models are available in the code. All backbone checkpoints will be made publicly available upon paper acceptance.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017. URL https://arxiv.org/abs/1610.01644.

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 3rd edition, 2003.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255. PMLR, 2013. URL https://proceedings.mlr.press/v28/andrew13.html.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://arxiv.org/pdf/2106.07682.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2105.04906.

Aleksandar Botev, Andrew Jaegle, Peter Wirnsberger, Daniel Hennes, and Irina Higgins. Which priors matter? Benchmarking models for learning latent dynamics. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/pdf/2111.05458.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. doi: 10.1017/CBO9780511804441.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020. URL https://arxiv.org/abs/2003.04630. ICLR 2020 Deep Differential Equations Workshop.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2303.09166.

Peter Gács and János Körner. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2(2):149–162, 1973.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

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 *Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*. PMLR, 2019. URL https://arxiv.org/pdf/1905.06642.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.14486.

Sharut Gupta, Sanyam Kansal, Stefanie Jegelka, Phillip Isola, and Vikas Garg. Canonicalizing multimodal contrastive representation learning. In *ICLR 2026 Workshop on Representational Alignment (Re-Align)*, 2026. URL https://arxiv.org/abs/2602.17584.

Souhail Hadgi, Luca Moschella, Andrea Santilli, Diego Gomez, Qixing Huang, Emanuele Rodolà, Simone Melzi, and Maks Ovsjanikov. Escaping Plato's cave: Towards the alignment of 3D and text latent spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 19825–19835, 2025. URL https://arxiv.org/pdf/2503.05283.

Irina Higgins, Peter Wirnsberger, Andrew Jaegle, and Aleksandar Botev. SyMetric: Measuring the quality of learnt Hamiltonian dynamics inferred from vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 25591–25605, 2021. URL https://arxiv.org/pdf/2111.05986.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/pdf/2006.11239.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. doi: 10.1093/biomet/28.3-4.321.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2405.07987.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. URL https://arxiv.org/pdf/1605.06336.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2505.12540.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL https://arxiv.org/abs/2102.05918.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://arxiv.org/pdf/1907.04809.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into Plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026. URL https://arxiv.org/abs/2604.18572.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://arxiv.org/pdf/1905.00414.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004. doi: 10.1103/PhysRevE.69.066138. URL https://arxiv.org/abs/cond-mat/0305641.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2: 4, 2008. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full.

Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. VideoFlow: A conditional flow-based model for stochastic video generation. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1903.01434.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 2024. URL https://arxiv.org/abs/2310.16834.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research (TMLR)*, 2025. URL https://arxiv.org/abs/2401.03048.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Mohamed El Amine Seddik, Karttikeya Mangalam, and Noel E. O'Connor. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14334–14343, 2024. URL https://arxiv.org/pdf/2401.05224.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/pdf/2209.15162.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/pdf/2209.15430.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. ASIF: Coupled data turns unimodal models to multimodal without training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/pdf/2210.01738.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2212.09748.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://arxiv.org/abs/2103.00020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E. Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. Cross-entropy is all you need to invert the data generating process. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.21869.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3–4):441–451, 1959. doi: 10.1007/BF02024507.

Geoffrey Roeder, Luke Metz, and Diederik P. Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. URL https://arxiv.org/pdf/2007.00810.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2112.10752.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.07524.

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! Towards vision–language correspondence without parallel data. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 24983–24992, 2025. URL https://arxiv.org/pdf/2503.24129.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022. URL https://arxiv.org/abs/2210.08402.

Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.16282.

Antonio Torralba and Yair Weiss. A theory of contrastive learning with natural images. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/pdf/2607.07470.

Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian generative networks. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/pdf/1909.13789.

Pantelis Vafidis, Aman Bhargava, and Antonio Rangel. Disentangling representations through multitask learning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2407.11249.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018. doi: 10.1017/9781108231596.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://arxiv.org/pdf/2106.04619.

H. S. Witsenhausen. On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113, 1975. doi: 10.1137/0128010.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL https://arxiv.org/abs/2103.03230.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference (MLHC)*, volume 182 of *Proceedings of Machine Learning Research*. PMLR, 2022. URL https://arxiv.org/abs/2010.00747.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ODE-Net: Learning Hamiltonian dynamics with control. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1909.12077.

Tyler Zhu, Tengda Han, Leonidas Guibas, Viorica Pătrăucean, and Maks Ovsjanikov. Dynamic reflections: Probing video representations with text alignment. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2511.02767.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 2021. URL https://arxiv.org/pdf/2102.08850.

Liu Ziyin and Isaac Chuang. Proof of a perfect Platonic representation hypothesis. *arXiv preprint arXiv:2507.01098*, 2025. URL https://arxiv.org/abs/2507.01098.

Liu Ziyin, Isaac Chuang, Tomer Galanti, and Tomaso Poggio. Formation of representations in neural networks. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.03006.

APPENDIX

We provide detailed results together with implementation details and a fuller theoretical treatment.

The appendix is organized as follows.

- Section A describes the dataset and systems, encompassing:
  - The Hamiltonian of each system (Appendix A.1)
  - Sampling ranges, the caption ladder, what each caption level discloses, templates and example captions (Appendix A.2)
- Section B gives model and training details: encoders and read-outs, model sizes, the auxiliary head and its loss balancing, the multi-view baselines, map-structure statistics, seeds and compute.
- Section C defines every metric, the ceiling-normalized CKA and its ceilings, the evaluation pools and their chance levels, and each metric's invariance class.
- Section D gives the full results on the dissociation (§6.1), encompassing:
  - Per-system alignment and decodability (Appendix D.1)
  - Permutation-calibrated metrics and what the no-reference floor contains (Appendix D.2)
  - Decodability under the enriched captions and the full-latent probe (Appendix D.3)
  - The information each representation carries and the information the two share (Appendix D.4)
- Section E covers the cross-modal maps (§6.2), encompassing:
  - Map classes, the tuned penalty, both retrieval directions, the round trip, physics transfer and the symmetry tests (Appendix E.1)
  - A nonlinear map search at $\lambda{=}0$, including an invertible flow in both directions (Appendix E.2)
  - Post-hoc alignment from paired anchors (Appendix E.3)
  - Neighbourhood size, intrinsic dimension and map orthogonality under whitening (Appendix E.4)
- Section F covers what the target must carry (§6.3), encompassing:
  - Target variants, the measured H-subspace dimension, minimal and partially overlapping targets, the phase head and the loss balance (Appendix F.1)
  - The auxiliary head reduced to a single affine map (Appendix F.2)
- Section G gives the robustness and scaling results (§6.4), encompassing:
  - Read-out compression (Appendix G.1)
  - Text read-out and caption length (Appendix G.2)
  - A different vision backbone (Appendix G.3)
  - Different text objectives (Appendix G.4)
  - Backbone scale, data scale and training breadth (Appendix G.5)
  - Zero-shot pretrained pairs (Appendix G.6)
- Section H analyses the held-out systems for (§6.4):
  - Where the gap survives, the rank of the learned map and qualitative transfer, followed by per-system metrics (Appendix H.1)
  - Three-body system by dynamical regime (Appendix H.2).
- Section I contains the theory, encompassing:
  - Proofs for §4 (Appendix I.1)
  - Why marginal objectives leave the frame free (Appendix I.2)
  - The low-dimensional shared subspace (Appendix I.3)
  - Removing the relative frame freedom, and the asymmetric-target case (Appendix I.4)

A Dataset and Systems

Dataset. Our data builds on DeepMind's Hamiltonian Dynamics Suite (Botev et al., 2021), which we extend with new system families, a charged (Coulomb) system, and the natural-language caption ladder. Training uses **eight base datasets**: four systems (mass–spring, pendulum, double pendulum and two-body gravity), each in a *conservative* and a *damped* (friction, η=0.05) variant. We additionally generate **two out-of-distribution datasets**, three-body and charged two-body (Coulomb), held out entirely from training. These two probe complementary axes of generalization: three-body is OOD in the *number of interacting bodies* (and is chaotic), while the charged system is OOD in the *force law* itself, a sign-varying Coulomb interaction that, unlike gravity, can either pull the bodies together or push them apart depending on the sign of the charges. Per system we render 7500 trajectories (6000 train / 1500 test) at $\Delta t = 0.05$ over 61 frames. Altogether the eight training datasets total 60,000 samples. Trajectories are integrated with `scipy.integrate.solve_ivp` (RK45, Dormand–Prince 5(4), adaptive step, $\mathrm{rtol} = \mathrm{atol} = 10^{-9}$), the damped variants adding $\dot{p} \mathrel{-}= \eta\dot{q}$ to the Hamiltonian field; the conservative mass–spring uses its closed-form solution. Solutions are output on a 0.005 grid and every tenth state kept, giving 61 frames at $\Delta t = 0.05$; the encoders see 16 of these at stride 4.

A.1 Equations

The Hamiltonian governing each system is given in Table 3.

Table 3: Dynamical systems, their Hamiltonians $H(q,p)$, dynamical regime, and which quantities vary across trajectories (masses m_i, the force constant, charges c_i); $r_{ij} = \|q_i - q_j\|$. All systems except charged two-body are generated in a conservative and a damped (friction, η=0.05) variant. $\star$ = held out from training, used only for out-of-distribution evaluation.

System	$H(q,p)$	Regime	Masses	Force const.
Mass–spring	$\frac{p^2}{2m} + \frac{1}{2}kq^2$	integrable, 1 DOF	✓	✓ (k)
Pendulum	$\frac{p^2}{2m\ell^2} + mg\ell(1-\cos q)$	integrable, 1 DOF	✓	✓ (g)
Double pendulum	H_{dp} (above)	chaotic, 2 DOF	✓	✓ (g)
Two-body (gravity)	$\sum_i \frac{\|p_i\|^2}{2m_i} - \frac{Gm_1m_2}{r_{12}}$	integrable (Kepler)	✓	✓ (G)
Three-body $\star$	$\sum_i \frac{\|p_i\|^2}{2m_i} - G\sum_{i<j} \frac{m_im_j}{r_{ij}}$	chaotic	✓	✓ (G)
Charged two-body $\star$	$\sum_i \frac{\|p_i\|^2}{2m_i} + \frac{kc_1c_2}{r_{12}}$	integrable, sign-varying	✓	✓ (k)

$$H_{\mathrm{dp}} = \frac{m_2\ell_2^2\,p_1^2 + (m_1{+}m_2)\ell_1^2\,p_2^2 - 2m_2\ell_1\ell_2\,p_1p_2\cos(q_1{-}q_2)}{2\,m_2\ell_1^2\ell_2^2\,[\,m_1 + m_2\sin^2(q_1{-}q_2)\,]} - (m_1{+}m_2)\,g\,\ell_1\cos q_1 - m_2\,g\,\ell_2\cos q_2$$

A.2 Generation Parameters

Table 4 lists the sampling range of every parameter. All are drawn uniformly and independently per body and per trajectory; the force constant is disclosed only in some L1/L2 captions, and the energy is never disclosed, so it must be inferred.

Captions. Captions are generated from deterministic templates (seeded by trajectory index) whose wording and vocabulary we authored with an LLM, specifying exactly which observations each template may disclose. The wording is *not* one fixed sentence per system: for each trajectory an

Table 4: Sampling ranges per system, drawn uniformly and independently per body and per trajectory. "Init. radius" is the initial position or orbit magnitude; for the charged system the charge magnitude is sampled and the sign pattern fixed per variant ($++/+-/--$).

System	Mass m	Force const.	Length ℓ	Charge $\lvert c\rvert$	Init. radius
Mass–spring	$[0.1, 3.0]$	$k \in [0.3, 12.0]$	–	–	$[0.2, 1.8]$
Pendulum	$[0.2, 2.0]$	$g \in [1.0, 9.0]$	$[0.4, 2.0]$	–	$[1.0, 2.5]$
Double pendulum	$[0.3, 1.1]$	$g \in [1.5, 6.0]$	$[0.5, 1.5]$	–	$[1.0, 2.5]$
Two-body (gravity)	$[0.3, 2.2]$	$G \in [0.5, 2.0]$	–	–	$[0.5, 2.0]$
Three-body	$[0.3, 2.2]$	$G \in [0.5, 2.0]$	–	–	$[0.5, 1.6]$
Charged two-body	$[0.5, 2.0]$	$k \in [0.5, 2.0]$	–	$[0.5, 2.0]$	$[1.0, 2.5]$

index-seeded RNG assembles the caption from a large space of interchangeable phrasings: intros, qualitative frames, verbs, clause orderings, and which of the visible clues to name. So different trajectories of the same system receive different sentences and no fixed template can be memorized; so the encoder must read the disclosed content. The text discloses only *raw observations* (sampled positions and momenta) and the visible masses and lengths, read off the rendered disc sizes and swing radius; it never quotes the energy and quotes the force constant only in some L1/L2 captions, so recovering the physics is largely an inference task. A four-level ladder L0–L3 varies *which* raw observations are disclosed, not a strictly nested amount , while holding appearance (color and object count) fixed across the ladder:

- **L0**: a one-line qualitative description of the motion + a few *interleaved* phase-space samples (position *and* momentum at each step; the double pendulum is the lone exception, with positions only); lengths may be stated (only double pendulum).
- **L1**: a denser set of *position-only* samples + the visible masses (from disc sizes); no momenta.
- **L2**: masses + a fuller description + interleaved phase-space (or denser position) samples, depending on the system.
- **L3**: the fullest description + masses + interleaved phase-space samples (position *and* momentum).

What is and is not disclosed. Visible *appearance* (color, count) is stated at every level. The *masses*, read from disc sizes, enter at L1, as do the *lengths*, read from the swing radius (except for the double pendulum); for the charged (evaluation-only) system the *charges* are also stated at L3. The *energy* is the only physical quantity never disclosed, at any level or for any system. The *force constant* is withheld at L0 and L3 everywhere, and never disclosed for the pendulum and double pendulum. For the remaining systems it belongs to an equal-difficulty anchor set from which L1 and L2 each state one member, so it appears at one of these levels in a fraction of trajectories: always for mass–spring, whose set is $\{m, k\}$, and otherwise with the probability that the per-trajectory shuffle selects it.

Caption template (schema, pendulum). Slots in $\langle\cdot\rangle$ are filled per trajectory; the gravitational constant g and the energy never appear. Phase-space samples *interleave* θ and p at each timestep, at a uniform stride stated once (not per point); the sample budget is chosen so the tokenized caption fits the 256-token window.

L0: ⟨motion description⟩. arm length ℓ $\sim$ ⟨·⟩ | phase-space samples (position and momentum), one every ⟨Δt⟩s in order: theta=⟨·⟩ rad, p=⟨·⟩; ... Rendered as one ⟨color⟩ disc.

L1: arm length $\ell \sim$ ⟨·⟩, mass $m \sim$ ⟨·⟩, positions, one every ⟨$\Delta t'$⟩s in order: theta=⟨·⟩ rad; ... (denser, angle-only).

L2: ⟨fuller description⟩; mass m $\sim$ ⟨·⟩, arm length ℓ $\sim$ ⟨·⟩, phase-space samples (position and momentum), one every ⟨Δt⟩s in order: theta=⟨·⟩ rad, p=⟨·⟩; ...

L3: ⟨fullest description⟩; mass m $\sim$ ⟨·⟩, arm length ℓ $\sim$ ⟨·⟩,

`phase-space samples (position and momentum), one every` $\langle \Delta t \rangle$`s in order: theta=`$\langle \cdot \rangle$` rad, p=`$\langle \cdot \rangle$`; ...`

Each trajectory carries one caption per level (four total, L0–L3), a single deterministic wording seeded by the trajectory index; phrasing varies across *trajectories*, systems, and levels through the seeded templated composition, giving $240{,}000$ captions in total ($192{,}000$ train).

Example captions (L0–L3). Real captions for one pendulum (single-body) trajectory, verbatim from the default set. The state is the angle θ and its momentum p; phase-space samples *interleave* θ and p at each timestep, listed at a uniform stride stated once ("one every $\langle \Delta t \rangle$s in order"). The disclosed observations vary by level (interleaved (θ, p) at L0/L2/L3, denser angle-only samples at L1), while the gravitational constant g and the energy are *never* stated.

> **L0.** A swinging bob, swinging through a wide arc. It sweeps through an arc and swings back. Pinned down: the swing radius shows an arm length of about $\ell \sim 1.33$ | phase-space samples (position and momentum), one every 0.4s in order: theta=1.11 rad, p=-1.07; theta=0.03 rad, p=-1.7; theta=-1.07 rad, p=-1.12; theta=-1.48 rad, p=-0.0; theta=-1.07 rad, p=1.12; theta=0.02 rad, p=1.7; theta=1.1 rad, p=1.08; theta=1.48 rad, p=-0.05. Rendered as one purple disc.
>
> **L1.** A swinging bob … it moves quickest at the lowest point of its path. Ascertained, namely the swing radius shows an arm length of about $\ell \sim 1.33$, the bob's size shows its mass is about $m \sim 0.31$, positions, one every 0.25s in order: theta=1.11 rad; theta=0.49 rad; theta=-0.28 rad; theta=-0.96 rad; theta=-1.38 rad; theta=-1.48 rad; theta=-1.25 rad; theta=-0.72 rad; theta=0.02 rad; theta=0.76 rad; theta=1.27 rad; theta=1.48 rad. Rendered as one purple disc.
>
> **L2.** A long arm carrying a light bob … it moves quickest at the lowest point of its path; it slows as it climbs toward the ends of the swing. Gathered, namely the bob's size shows its mass is about $m \sim 0.31$, the swing radius shows an arm length of about $\ell \sim 1.33$, phase-space samples (position and momentum), one every 0.4s in order: theta=1.11 rad, p=-1.07; …; theta=1.48 rad, p=-0.05. The scene shows a single purple circle.
>
> **L3.** A swinging bob … the arc turns around at each extreme. Extracted: the bob's size shows its mass is about $m \sim 0.31$, the swing radius shows an arm length of about $\ell \sim 1.33$, phase-space samples (position and momentum), one every 0.4s in order: theta=1.11 rad, p=-1.07; …; theta=1.48 rad, p=-0.05. A lone purple circle is depicted.

Appearance / colors. Bodies are drawn as colored discs whose radius encodes mass ($r \propto \sqrt{m}$). Colors are eight evenly-spaced hues, named in the captions as *red, orange, yellow, green, cyan, blue, purple,* and *magenta*. Appearance is an extra nuisance factor the generative/autoregressive objective must model but the physics does not depend on, and it is held constant across the L0–L3 ladder.

B Model and Training Details

Encoders and defaults. The vision encoder is a Latte (Ma et al., 2025) diffusion transformer (ϵ-prediction, linear schedule, 1000 steps, patch 4 giving 1024 tokens per 16-frame clip) and the text encoder a GPT-2 transformer (Radford et al., 2019), width-matched so both latent spaces share dimension d. The two share no weights and are trained separately, one per GPU, each with its own optimizer (Table 5). The video encoder consumes $F{=}16$ frames at $32{\times}32$ scaled to $[-1, 1]$; the text encoder consumes captions tokenized by a caption-specific word tokenizer: text is lowercased and split into letter runs and single non-letter characters, so numbers are read digit by digit; the vocabulary is built over all training captions (353 tokens) and captions are capped at 256 tokens.

Read-out. Each backbone produces a sequence of hidden states, $H_{\mathrm{V}}(x) \in \mathbb{R}^{L_{\mathrm{V}} \times d}$ (over frames and patches) and $H_{\mathrm{T}}(y) \in \mathbb{R}^{L_{\mathrm{T}} \times d}$ (over tokens). The representation used for all downstream analysis is a fixed read-out $r : \mathbb{R}^{L \times d} \to \mathbb{R}^d$ of this sequence,

$$f(x) = r\big(H_{\mathrm{V}}(x)\big), \qquad g(y) = r\big(H_{\mathrm{T}}(y)\big), \tag{8}$$

in one of two forms: a mean pool,

$$r_{\mathrm{pool}}(H) = \tfrac{1}{L} \textstyle\sum_{l=1}^{L} H_l, \tag{9}$$

Table 5: Default optimization per encoder. A held-out 5% validation split selects the best checkpoint.

	Vision (Latte)	Text (GPT-2)
Optimizer	AdamW	AdamW
Learning rate	1×10^{-4}	2.5×10^{-4}
Weight decay	0	0.01
Schedule	constant	warmup $\rightarrow$ cosine
Epochs	30	30
Batch size	4	4

or a learned K-query bottleneck, in which K learned queries $Q \in \mathbb{R}^{K\times d}$ cross-attend to the sequence,

$$r_{\mathrm{bn}}(H) = \mathrm{vec}\big(\mathrm{ATTN}(Q, H, H)\big) \in \mathbb{R}^{Kd}, \tag{10}$$

where ATTN is a 4-head cross-attention block with learned projections and the K output rows are concatenated, so the bottleneck latent has dimension Kd ($768/1536/3072/6144$ for $K = 2/4/8/16$). The auxiliary head reads the same vector; generation is conditioned on it through a separate one-query attention.

Scale. We train three matched sizes, S/B/L, scaling the Latte backbone with the GPT-2 width, heads and layers matched at each step; Table 6.

Table 6: Model sizes. "Latte-X/4" is a patch-4 Latte preset; d is the shared latent width. Parameter counts are per encoder and include the token and position embeddings, the tied output embedding, and the phase head; the auxiliary head and conditioning pool are excluded.

Size	Vision (hidden/depth/heads)	Text (embd/layer/head)	d	Vision params	Text params
S	Latte-S/4 (384/12/6)	384/6/6	384	32.7M	11.6M
B	Latte-B/4 (768/12/12)	768/12/12	768	130.2M	88.88M
L	Latte-L/4 (1024/24/16)	1024/24/16	1024	456.9M	308.71M

Auxiliary head. A light head ρ cross-attends to the encoder tokens and regresses the ground-truth h. The auxiliary loss is a weighted sum over groups,

$$\mathcal{L}_{\mathrm{aux}}(z) = w_{\mathrm{par}}\mathcal{L}_{\mathrm{params}} + w_{\mathrm{en}}\mathcal{L}_{\mathrm{energy}} + w_{\mathrm{sys}}\mathrm{CE}(\mathrm{system})\ (+\,\mathcal{L}_{\mathrm{phase}}),$$

with the components in Table 7 (the phase term is active only when the phase head is enabled). The parameter head is *masked*: only the entries active for a given system contribute. Setting $\lambda{=}0$ removes the head entirely.

Table 7: Auxiliary-head groups. Regression uses a Huber (smooth-L_1) loss; the parameter group masks to the active entries per system.

Group	Target	Loss	Weight
params	Hamiltonian parameters θ (masked)	masked Huber	1.0
energy	total energy (a derived scalar)	Huber	0.5
system	system identity	cross-entropy	1.0
phase	per-frame state $(q, p)(t)$	Huber	1.0

Auxiliary loss: normalization and adaptive weighting. Because the terms live on very different scales (a diffusion reconstruction, a Huber on normalized parameters, a cross-entropy on the system

label), we normalize each term by an exponential-moving-average (EMA, momentum 0.99) of its own *detached* magnitude before weighting, so every term contributes at unit scale:

$$\widehat{\mathcal{L}}_i = \frac{\mathcal{L}_i}{\mathrm{EMA}(|\mathcal{L}_i|) + \epsilon}, \qquad \epsilon = 10^{-6}. \tag{11}$$

The auxiliary loss is the weighted sum of the normalized head losses,

$$\mathcal{L}_{\mathrm{aux}} = \lambda_{\mathrm{aux}} \sum_g w_g \, \widehat{\mathcal{L}}_g, \qquad w = \{\text{params:1.0, energy:0.5, system:1.0}\}, \tag{12}$$

with the phase head added at weight 1. The global gate λ_{aux} turns the branch on (λ_{aux}=1) or off (λ_{aux}=0, our no-aux control). Reconstruction and auxiliary losses are combined by an *adaptive* balancer:

$$\mathcal{L} = \gamma \, \widehat{\mathcal{L}}_{\mathrm{recon}} + (1-\gamma)\, \mathcal{L}_{\mathrm{aux}}, \qquad \gamma = \frac{\mathrm{EMA}(|\widehat{\mathcal{L}}_{\mathrm{recon}}|)}{\mathrm{EMA}(|\widehat{\mathcal{L}}_{\mathrm{recon}}|) + \mathrm{EMA}(|\mathcal{L}_{\mathrm{aux}}|)}. \tag{13}$$

In terms of Eq. (4), the gate and balancer implement $\lambda = 1-\gamma$ when $\lambda_{\mathrm{aux}} = 1$ and $\lambda = 0$ when $\lambda_{\mathrm{aux}} = 0$; the Pareto statement of §4 therefore covers every $\gamma \in (0,1)$ of Table 31. Here γ is the fraction of the gradient budget spent on reconstruction and $(1-\gamma)$ the fraction on the auxiliary target; γ is detached, so it reweights without contributing its own gradient. Both terms are unit-scale, so the equilibrium is $\gamma \approx 1/(1+\sum_g w_g)$, well below parity: $\gamma \approx 0.22$ for both encoders. The fixed-γ sweep (Table 31) shows the balance is immaterial.

Training. All models train for 30 epochs at batch size 4 on 16-frame clips, with a 5% validation split for the best checkpoint. The manipulated variable is λ: λ>0 enables the auxiliary head, λ=0 removes it, with data, architecture, and optimization otherwise fixed. Figure 5 shows the training and validation loss curves for the Latte-S mains and the K-query bottleneck runs.

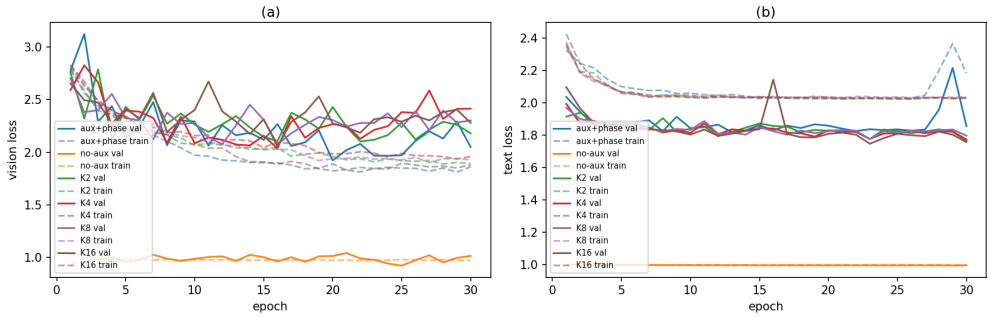

Figure 5: **Training and validation loss for the small (Latte-S) models.** Vision (a) and text (b) train/validation loss vs epoch for the aux+phase and no-auxiliary mains and the K-query bottleneck runs (solid: validation; dashed: training). Validation loss flattens well before the 30-epoch budget, so the small models are trained to convergence.

Multi-view SSL baselines. VICReg (Bardes et al., 2022) and Barlow Twins (Zbontar et al., 2021) align the two encoders with an implicit view-matching loss and *no* ground-truth h. Both encoders train jointly on one GPU (the loss needs both latents in the same step) with expander heads of width 512 on the pooled read-outs (so the *evaluated* representation is identical to every other run; the expanders are discarded at evaluation), batch 16, the generative term retained. The baseline optimizes $\mathcal{L} = \mathcal{L}_{\mathrm{gen}} + \beta\,\mathrm{VICReg}(\psi_V(z_V), \psi_T(z_T))$, with per-modality expanders ψ_V, ψ_T (3-layer $\mathrm{Linear} \to \mathrm{BN} \to \mathrm{ReLU}$, width 512) and β=1.0 set so the SSL and generative terms are comparable in magnitude. The VICReg term is the standard invariance/variance/covariance sum (Bardes et al., 2022) with coefficients $(\lambda_{\mathrm{inv}}, \mu_{\mathrm{var}}, \nu_{\mathrm{cov}}) = (25, 25, 1)$; Barlow Twins uses off-diagonal weight $\lambda = 5\times10^{-3}$. Because these objectives match the two representations directly, they *couple* the encoders and align the spaces without any ground-truth target (CKA 0.54 / 0.68; Table 28), the coupling route of §6.3. The two objectives nonetheless differ at the *point* level: VICReg's invariance term minimizes

the distance between *paired* view-embeddings, pinning individual trajectories (mapped r@1 0.35), whereas Barlow Twins constrains only the feature cross-correlation matrix, a batch-level second-order statistic, toward identity, so it agrees on covariance without a per-trajectory correspondence (r@1 0.05, at the floor). As with CKA and retrieval, only a per-sample matching term installs point-level correspondence. Note that this baseline is trained jointly on one GPU with both latents visible in the same step, unlike the $\lambda{>}0/\lambda{=}0$ pair which train separately with no cross-modal signal at all; the comparison is therefore illustrative of the coupling route rather than a fully matched ablation. The alignment is not a batch-size or weighting artifact: a $4\times$-larger batch (64) with doubled invariance weight (50), all else fixed, aligns comparably (CKA 0.60, mutual-kNN 0.17, mapped r@1 0.61).

Map-structure statistics. For a fitted cross-modal map W we report four geometry summaries. The *round-trip deviation* $\mathrm{roundtrip_dev} = \|WW - I\|_F / \|I\|_F$ measures how far composing the two maps departs from identity (with retrieval back to the source). The *orthogonality deviation* $\|W^\top W / \|W\|_2^2 - I\|_F$ is 0 iff W is a scaled rotation. $\mathrm{corr}(W_{v\to t}, W_{t\to v}^\top) \to 1$ iff the two directions are transposes of one another, i.e. W is a rotation. Finally we report W's *effective rank* (the participation ratio $(\sum_i \sigma_i)^2 / \sum_i \sigma_i^2$ of its singular values), quantifying how many directions the map actually uses.

Seeds. The core dissociation, the target variants of Table 28, the out-of-distribution evaluation, the affine-head control, the fixed-γ sweep, the bottleneck sweep, and the post-hoc analyses (anchor sample-complexity, read-out sensitivity, mutual-kNN vs k, intrinsic dimensionality, map orthogonality) are each repeated over five seeds $\{7, 42, 123, 1337, 2024\}$ (init, data shuffle, and train/val split). The partial-overlap and 80k data-scaling ablations use five seeds $\{0, 42, 123, 1337, 2024\}$; the caption-length ($512$-token) ablation uses seeds $\{7, 42, 123\}$; the minimal-target, text-objective, vision-architecture, and auxiliary-configuration ablations and the mutual-information estimates of App. D.4 use a single seed (seed 0). Reported values are means (and, where shown, std) over the available seeds and the eight base systems.

Compute. Encoders train on a single GPU—an NVIDIA A100 (80 GB) or H200 (141 GB)—with 8 CPU cores; vision and text run as separate one-GPU jobs. All experiments use the Latte-S backbone ($d{=}384$) except the scale ablation (Table 40; Latte-S/B/L) and the vision-architecture ablation (Table 38; alternative backbones at matched $d{=}384$). For the Latte-S runs (8 base systems, 30 epochs) a single encoder takes 12–14 h on an H200 and 14–15 h on an A100; the modest gap reflects a small, largely data-loading-bound model, and wall-clock is essentially identical across the $\lambda{>}0$ and $\lambda{=}0$ conditions. The Latte-B scale runs take $\approx$29 h (vision) and $\approx$19 h (text), and the Latte-L scale point up to $\sim$4.5 days.

C Metric Definitions

Let $Z_V \in \mathbb{R}^{N\times d_V}$ and $Z_T \in \mathbb{R}^{N\times d_T}$ be the pooled latents for N paired trajectories, one vision vector and one text vector per trajectory. Unless stated otherwise both matrices are column-centered.

Table 8: Alignment metrics grouped by *level* (coarse global geometry versus point-level identity) and the transformations each is invariant to. SVCCA is affine-invariant (blind to any invertible linear frame); ceiling-normalized CKA is raw CKA rescaled by each space's self-reproducibility $\sqrt{c_V c_T}$.

metric	level	invariant to
CKA	global (coarse)	orthogonal + isotropic scale
RSA (Spearman)	global (rank)	monotone maps of distances
SVCCA	shared subspace	invertible linear (affine)
mutual-kNN	local / identity	none (embedding-level)
mapped r@1	point / identity	class of the fitted map

Linear CKA. Centered Kernel Alignment with a linear kernel, invariant to orthogonal transform and isotropic scaling:

$$\mathrm{CKA}(Z_V, Z_T) = \frac{\|Z_V^\top Z_T\|_F^2}{\|Z_V^\top Z_V\|_F \, \|Z_T^\top Z_T\|_F} \in [0, 1]. \tag{14}$$

Ceiling-normalized CKA: implementation. Cross-modal CKA is bounded in practice by how well an encoder reproduces itself across training seeds. For each system and condition, with $Z_V^{(s)}, Z_T^{(s)}$ the latents of seed s and $\mathcal{P}$ the 10 seed pairs $s < s'$ of the five seeds, the ceilings are

$$c_V = \frac{1}{|\mathcal{P}|} \sum_{(s,s') \in \mathcal{P}} \mathrm{CKA}\big(Z_V^{(s)}, Z_V^{(s')}\big), \qquad c_T = \frac{1}{|\mathcal{P}|} \sum_{(s,s') \in \mathcal{P}} \mathrm{CKA}\big(Z_T^{(s)}, Z_T^{(s')}\big), \tag{15}$$

and normalized CKA is $\mathrm{CKA}(Z_V, Z_T)/\sqrt{c_V c_T}$, each condition by its own ceilings. The ceilings use the same vectors as the cross-modal CKA: the mean-pooled read-out, the full paired test pool of each system (N=1500 trajectories, same order for every matrix), column-standardized, and the linear CKA of Eq. (14). Table 9 lists them; the normalized columns of Tables 10, 12 and 28 divide the reported raw CKA (and its $\pm$) by $\sqrt{c_V c_T}$: per system in Table 10, per regime in Table 12 (the held-out ceilings are computed the same way on each held-out system's test pool), and with the mean ceilings of each target in Table 28. Both encoders reproduce themselves well with and without the auxiliary target ($c_V/c_T = 0.90/0.96$ vs. $0.88/0.93$), so the λ=0 floor is not seed noise: normalized, it is 0.23 against 0.77.

Table 9: Self-reproducibility ceilings c_V, c_T: mean linear CKA between differently seeded copies of the same encoder, $\pm$ over seed pairs. The mean is over the eight base systems; the held-out rows give the ceilings behind the normalized column of Table 12.

	aux		no-aux	
System	c_V	c_T	c_V	c_T
Mass–spring	0.93±0.01	0.98±0.01	0.92±0.03	0.96±0.01
+ friction	0.94±0.01	0.98±0.01	0.92±0.03	0.96±0.01
Pendulum	0.91±0.01	0.96±0.01	0.89±0.03	0.95±0.01
+ friction	0.91±0.02	0.96±0.01	0.89±0.03	0.95±0.01
Double pendulum	0.83±0.03	0.93±0.02	0.86±0.03	0.88±0.02
+ friction	0.84±0.03	0.93±0.02	0.86±0.02	0.88±0.02
Two-body	0.93±0.02	0.97±0.01	0.86±0.02	0.92±0.01
+ friction	0.93±0.02	0.97±0.01	0.86±0.02	0.92±0.01
Mean	0.90±0.01	0.96±0.01	0.88±0.02	0.93±0.01
Held out (zero-shot)				
Three-body	0.93±0.02	0.87±0.04	0.88±0.05	0.89±0.01
+ friction	0.93±0.02	0.87±0.04	0.89±0.04	0.89±0.01
Charged two-body	0.81±0.03	0.85±0.06	0.87±0.03	0.85±0.02

Mutual k-nearest-neighbors. The average overlap between each point's neighborhood in vision space and in text space:

$$\mathrm{mkNN@}k = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathcal{N}_k^V(i) \cap \mathcal{N}_k^T(i)|}{k}, \tag{16}$$

with chance $\approx k/N$ and k=10; the local, identity-level convergence measure used by the Platonic Representation Hypothesis (Huh et al., 2024). Map-free.

Cross-modal retrieval (recall-at-1). We fit the map $W_{t\to v}$ (main text) and, for each caption embedding $g(y_i)$, rank all N vision latents by Euclidean distance to $W_{t\to v}\, g(y_i)$; a hit is scored when the true partner $z_{V,i}$ is the single nearest neighbor:

$$\mathrm{R@1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\arg\min_j \left\| W_{t\to v}\, g(y_i) - z_{V,j} \right\| = i\right], \tag{17}$$

chance $1/N$.

Evaluation pools (why the chance levels differ). The map-free measures (CKA, mutual-kNN, RSA, SVCCA) are computed on the full paired test pool, whereas the map-based measures (retrieval, Procrustes, transfer) are scored on the 30% held-out split used to fit the map. The two pools therefore have different N: mutual-kNN chance is $k/N \approx 0.007$ on the full pool, while retrieval chance is $1/N \approx 0.002$ on the smaller held-out split ($N \approx 450$). The two chance figures quoted in the main text reflect these two pools, not an inconsistency.

RSA. Form the within-modal similarity matrices S_V, S_T with $S_V[i,j] = \cos(z_{V,i}, z_{V,j})$ and take the Spearman rank correlation between their upper-triangular entries:

$$\mathrm{RSA} = \rho_{\mathrm{Spearman}}\big(\mathrm{vec}_{\triangle} S_V,\ \mathrm{vec}_{\triangle} S_T\big), \tag{18}$$

invariant to any similarity-order-preserving map. Map-free.

SVCCA. Project each space onto its top 99%-variance singular directions ($\tilde{Z}_V, \tilde{Z}_T$), then average the canonical correlations:

$$\mathrm{SVCCA} = \frac{1}{c}\sum_{i=1}^{c} \rho_i, \qquad \{\rho_i\} = \mathrm{CCA}(\tilde{Z}_V, \tilde{Z}_T). \tag{19}$$

SVCCA averages all canonical correlations, and the many moderate ones from content both latents carry keep it high in both conditions, so it does not dissociate them; we report it only for completeness.

Procrustes R^2. Fit the best orthogonal map $R^\star = \arg\min_{R^\top R = I} \|Z_T R - Z_V\|_F^2$ (SVD $Z_V^\top Z_T = U\Sigma V^\top$, $R^\star = VU^\top$) and report

$$R^2_{\mathrm{Proc}} = 1 - \frac{\|Z_V - Z_T R^\star\|_F^2}{\|Z_V - \bar{Z}_V\|_F^2}. \tag{20}$$

Invariance and why the measures differ. The measures differ in the class of transforms they ignore: RSA is invariant to any similarity-order-preserving map; CKA and Procrustes to orthogonal transforms (plus isotropic scaling for CKA); SVCCA to any invertible affine map. Without the shared reference the identity-level measures (retrieval, mutual-kNN) collapse to near chance and CKA and Procrustes fall to a low floor, whereas SVCCA does not dissociate, evidence that the shared reference induces a specific point-level correspondence, beyond a general linear relationship between the spaces. In the terms of Eq. (5), the transform class $\mathcal{T}$ is the general invertible linear maps: CKA and Procrustes certify its orthogonal-plus-scale sub-case, retrieval and transfer test the full class through a fitted T, and SVCCA shares T's invariance but averages over all canonical correlations; what dissociates the conditions is the number of near-unit ones (Table 32).

D The Dissociation

This appendix gives the full results for §6.1.

D.1 Per-system alignment

Table 10 reports the per-system alignment behind the Mean row quoted in §6.1

Table 10: Per-system cross-modal alignment, mean ± std over seeds, auxiliary ($\lambda>0$) vs no-auxiliary ($\lambda=0$); the Mean row gives mean ± std over the eight systems. Retrieval is text→vision, chance ≈ 0.002; ceilings for the normalized column are in Table 9.

	CKA		norm. CKA		mutual-kNN		r@1	
System	aux	no-aux	aux	no-aux	aux	no-aux	aux	no-aux
Mass–spring	0.76±0.02	0.20±0.03	0.80±0.02	0.21±0.03	0.47±0.03	0.05±0.01	0.85±0.03	0.03±0.01
+ friction	0.78±0.02	0.23±0.04	0.82±0.02	0.24±0.04	0.46±0.03	0.05±0.00	0.84±0.03	0.03±0.01
Pendulum	0.66±0.02	0.17±0.03	0.71±0.02	0.18±0.03	0.39±0.01	0.04±0.00	0.90±0.02	0.03±0.01
+ friction	0.68±0.02	0.18±0.04	0.72±0.02	0.20±0.04	0.39±0.01	0.05±0.01	0.90±0.03	0.03±0.01
Double pendulum	0.60±0.02	0.17±0.05	0.68±0.02	0.20±0.06	0.32±0.02	0.02±0.00	0.89±0.03	0.03±0.01
+ friction	0.61±0.03	0.16±0.05	0.69±0.03	0.18±0.06	0.32±0.02	0.02±0.01	0.91±0.03	0.04±0.01
Two-body	0.84±0.00	0.28±0.07	0.88±0.00	0.31±0.08	0.46±0.09	0.07±0.02	0.73±0.12	0.08±0.02
+ friction	0.83±0.01	0.30±0.06	0.87±0.01	0.34±0.07	0.45±0.09	0.07±0.02	0.73±0.11	0.07±0.02
Mean	**0.72±0.09**	**0.21±0.07**	**0.77±0.10**	**0.23±0.08**	**0.41±0.08**	**0.047±0.019**	**0.84±0.09**	**0.042±0.023**

Table 11: Parameter decodability. Held-out linear-probe R^2 for the system parameters recovered from each *frozen* encoder (mean ± std over seeds and the eight base systems), with and without the shared reference.

	vision params R^2	text params R^2
no-aux ($\lambda=0$)	0.79±0.13	0.85±0.11
aux ($\lambda>0$)	0.96±0.04	0.95±0.03

Table 12: Core cross-modal dissociation, aux ($\lambda>0$) vs. no-aux ($\lambda=0$): mean over the eight in-distribution systems and zero-shot on the two held-out systems, mean ± std over seeds. Per-system values in Tables 10 and 47.

	CKA		norm. CKA		kNN		r@1	
regime	aux	no	aux	no	aux	no	aux	no
In-dist. (8 sys)	0.72±0.09	0.21±0.07	0.77±0.10	0.23±0.08	0.41±0.08	0.047±0.019	0.84±0.09	0.042±0.023
OOD 3-body	0.59±0.04	0.56±0.04	0.66±0.04	0.63±0.04	0.11±0.02	0.05±0.00	0.26±0.06	0.05±0.01
OOD Coulomb	0.41±0.04	0.22±0.06	0.49±0.05	0.25±0.07	0.19±0.03	0.03±0.00	0.49±0.11	0.02±0.01

D.2 Permutation-Calibrated Metrics

Each alignment metric is computed on the true caption–video pairing and on 200 random permutations of the pairing (a null that leaves both marginal spaces unchanged but destroys the correspondence). Observed CKA and RSA sit 70–260σ above the null (Table 13); the shuffled-pairing null CKA is ≈ 0.02 in both conditions, so neither the scores nor the aux-vs-no-aux gap are artifacts of width, depth, effective dimensionality, or sampling.

Table 13: Permutation-calibrated alignment (mean over eight systems): each metric on the true pairing vs. a shuffled-pairing null (200 permutations). z is in standard deviations of the null.

cond.	metric	observed	null	z
aux	CKA	0.73 ± 0.01	0.02	260
aux	RSA	0.67 ± 0.02	0.00	247
no-aux	CKA	0.23 ± 0.03	0.02	80
no-aux	RSA	0.20 ± 0.03	0.00	72

What the no-auxiliary floor contains. The floor is reliably above the permutation null, and reflects coarse similarity. It is a mean of per-system CKAs (Table 10), each system evaluated in isolation, so it is not cross-system identity or appearance clustering; it reflects physical content linearly shared by both spaces, which raises a covariance statistic such as CKA without tying individual trajectories together. The identity-level measures that would register convergence stay near chance (mutual-kNN 0.047, mapped r@1 0.042).

D.3 Enriched-Text Decodability

Table 14 gives the per-system decodability behind the competence check of §6.1: with the rich interleaved phase-space captions used throughout, the parameters are linearly decodable from the *no-auxiliary* text encoder on every system, mean $R^2 = 0.85$, while alignment stays at the floor (Table 10). Table 15 repeats the probe at three read-outs.

Table 14: Rich (interleaved phase-space) caption parameter decodability, per system: held-out probe R^2 from the *no-auxiliary* pooled text latent. Probe as in Table 11 (ridge penalty selected on an inner split); mean 0.85.

system	no-aux text params R^2
mass–spring	0.93 ± 0.01
+ friction	0.94 ± 0.01
pendulum	0.78 ± 0.02
+ friction	0.79 ± 0.02
double pendulum	0.70 ± 0.06
+ friction	0.72 ± 0.06
two-body	0.97 ± 0.01
+ friction	0.97 ± 0.01
mean	$\mathbf{0.85}\pm0.11$

D.4 Information in the Representations

We estimate three mutual informations per system on the dumped Latte-S latents: what each latent carries about the active Hamiltonian parameters, $I(z_v; h_{\mathrm{par}})$ and $I(z_t; h_{\mathrm{par}})$, and what the two share, $I(z_v; z_t)$. Here h_{par} is the per-system parameter block of h, z-scored. Two estimators are used, and they measure different things: the Kraskov–Stögbauer–Grassberger (KSG) estimator (Kraskov et al., 2004) (k=5) on PCA-8 latents, which bounds the information in eight directions, and the Gaussian

Table 15: Full-latent probe on the encoders trained without the reference (λ=0), mean $\pm$ std over 8 systems. Left: within-modal parameter decodability R^2. Right: cross-modal alignment (CKA, text$\to$vision r@1; chance 0.0017) at three read-outs.

read-out	text dec. R^2	vision dec. R^2	align CKA	align r@1
pooled	0.85±0.11	0.79±0.13	0.23±0.06	0.02±0.02
full (PCA)	0.40±0.19	0.64±0.21	0.15±0.03	0.03±0.02
h-subspace	n/a	n/a	0.53±0.12	0.09±0.10

estimate from the full held-out canonical spectrum, $I = -\frac{1}{2}\sum_i \log_2(1-\rho_i^2)$, the soft form of the near-unit count of Table 32. Permuting the rows of one variable sets the true MI to zero; both then return 0.00–0.06 bits.

The two conditions separate on shared information and not on parameter information (Figure 6). The latents share 31.8±6.0 bits with the reference against 9.4±2.8 without (4.31±0.95 against 1.16±0.41 by KSG), while both λ=0 encoders carry most of what the aux encoders carry about h_{par} (6.9±0.9 against 10.9±1.4 bits, vision) at a CKA that stays at the floor. The λ=0 latents carry the parameters; what the reference raises is the information the two latents share, to the level identification needs.

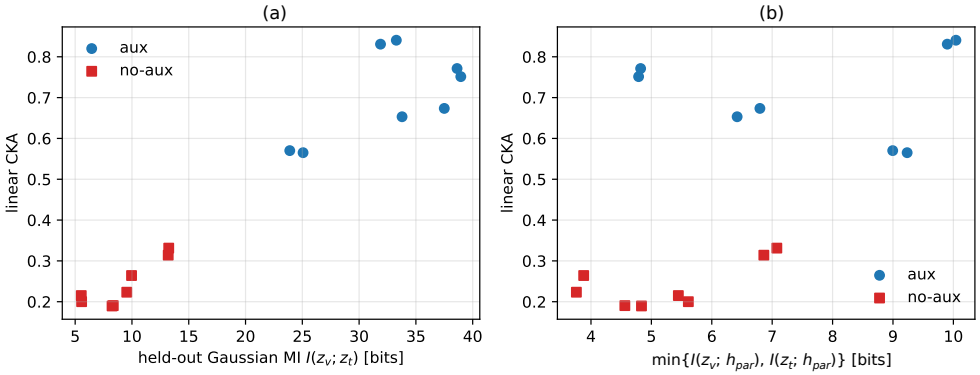

Figure 6: Alignment tracks shared information (8 systems, 1 seed). Circles: aux; squares: λ=0. Left: linear CKA vs held-out Gaussian $I(z_v; z_t)$. Right: CKA vs $\min\{I(z_v; h_{\mathrm{par}}), I(z_t; h_{\mathrm{par}})\}$, h_{par} the active Hamiltonian parameters.

E Cross-Modal Maps

This appendix gives the full results for §6.2.

E.1 Cross-modal maps and transfer

We report the cross-modal map three ways: map quality by class (Table 16), round-trip fidelity (Table 17), and within- vs. cross-modal physics transfer (Table 18). All read the mean-pooled latent of Eq. 9, the same representation for both conditions and for the native and mapped columns alike. The four map classes, least-squares, a trained linear layer, an MLP, and a transformer, are fit on the 70% split of the paired evaluation set and scored on the held-out 30%, so the reported R^2 and retrieval reflect generalization, not fit. The linear map is a single $d \times d$ layer; the MLP has two hidden layers with GELU; the transformer is a 2-layer, 4-head encoder over the latent tokens. Under the shared reference the richer maps add under three points over least-squares, and without it none recovers the correspondence (whole-latent $R^2 \leq 0.23$ even with a tuned penalty, and text$\to$vision r@1 near chance on held-out pairs), so the failure at λ=0 lies in the representations, rather than in the mapper's capacity. The fixed penalty (α=1) does under-regularize the λ=0 map: selecting α on an

inner split (median α: $1 \rightarrow 1000$) raises its R^2 from 0.00 to 0.23, yet leaves text→vision retrieval at the floor (0.03), and selecting α to maximize retrieval itself gives 0.04; with the reference, retrieval is 0.84 under every choice (Table 16, lower blocks), and the floor holds system by system (Table 19).

Table 16: Cross-modal map quality by map class (text→vision, mean over eight systems and seeds); best map class per column in bold. Lower blocks: ridge maps whose penalty α is selected on an inner split, for R^2 or for r@1 itself, in both directions; cosine is not recomputed there.

	aux			no-aux		
Map	R^2	cos	r@1	R^2	cos	r@1
Least-squares	0.77±0.08	0.87±0.05	**0.84±0.09**	0.00±0.10	0.36±0.07	0.04±0.02
Linear	0.76±0.08	0.86±0.04	0.79±0.09	**0.19±0.07**	**0.42±0.06**	0.04±0.02
MLP	**0.79±0.09**	**0.88±0.05**	**0.84±0.10**	−0.18±0.10	0.33±0.06	**0.05±0.02**
Transformer	**0.79±0.08**	**0.88±0.04**	**0.84±0.10**	−0.02±0.09	0.35±0.07	0.04±0.02
text→vision, ridge						
α=1 (as above)	0.77±0.02	–	0.84±0.05	0.00±0.06	–	0.04±0.01
α tuned for R^2	0.77±0.02	–	0.84±0.05	0.23±0.04	–	0.03±0.01
α tuned for r@1	0.77±0.02	–	0.84±0.05	0.11±0.04	–	0.04±0.01
vision→text, ridge						
α=1	0.85±0.01	–	0.91±0.03	0.37±0.02	–	0.22±0.02
α tuned for R^2	0.86±0.01	–	0.90±0.03	0.44±0.02	–	0.20±0.01
α tuned for r@1	0.86±0.01	–	0.91±0.03	0.40±0.02	–	0.22±0.02

Table 17: Round-trip (cycle) fidelity: mapped recall-at-1 (per-system values are mean±std over seeds; the Mean row averages across the 8 systems), and matrix deviation $\|WW - I\|/\|I\|$. Without the target the vision cycle collapses (mean 0.43 vs 0.98) while the text cycle only partly degrades (0.83 vs 1.00).

	Text cycle		Vision cycle		
System	aux	no-aux	aux	no-aux	$\frac{\|WW-I\|}{\|I\|}$
Mass–spring	1.00±0.00	0.70±0.03	0.98±0.01	0.25±0.07	1.19±0.05
+ friction	1.00±0.00	0.69±0.04	0.98±0.01	0.24±0.07	1.19±0.04
Pendulum	1.00±0.00	0.83±0.10	0.99±0.00	0.43±0.10	1.87±0.43
+ friction	1.00±0.00	0.85±0.08	1.00±0.00	0.41±0.12	1.79±0.47
Double pendulum	1.00±0.00	0.87±0.12	1.00±0.00	0.48±0.10	2.30±0.04
+ friction	1.00±0.00	0.87±0.10	1.00±0.00	0.48±0.11	2.25±0.06
Two-body	1.00±0.00	0.92±0.08	0.96±0.03	0.58±0.08	1.30±0.08
+ friction	0.99±0.01	0.93±0.06	0.97±0.03	0.57±0.08	1.29±0.08
Mean	1.00±0.00	0.83±0.08	0.98±0.01	0.43±0.12	1.65±0.44

Cross-system retrieval gallery. The recall-at-1 we report is within-system: each text query is matched against the vision gallery of its own system (chance $\approx 1/450$). To check that this reflects a genuine per-trajectory correspondence rather than within-system clustering, we pool the held-out vision of *all* eight systems into one gallery (3600 items, chance $\approx 1/3600$) and re-run mapped text→vision retrieval. The shared-reference recall barely moves ($0.86 \rightarrow 0.79$ on this evaluation pool) despite the $8\times$ harder gallery, while the no-auxiliary control stays at the floor ($0.044 \rightarrow 0.022$): the map recovers the *specific* paired trajectory even among cross-system distractors, so the retrieval signal reflects per-trajectory identity.

Retrieval direction. Text→vision is the harder direction and the one we report. Vision→text retrieval at λ=0 is 0.20±0.01 with a tuned penalty (0.22±0.02 at α=1; map R^2 0.44), about $90\times$ chance, against 0.91 with the reference. The asymmetry follows the gallery: at λ=0 the text latent is higher-dimensional and more decodable than the vision latent (Tables 24, 11), so nearest-neighbour

Table 18: Cross-modal physics transfer, mean over seeds. Held-out R^2 within each modality (Vis, Text) and across the fitted linear map ($v \to t$, $t \to v$; chance $R^2 \leq 0$); probes and maps are ridge regressions whose penalty is selected on an inner split.

	Auxiliary (λ>0)				No-auxiliary (λ=0)			
System	Vis	Text	$v \to t$	$t \to v$	Vis	Text	$v \to t$	$t \to v$
Mass–spring	0.99±0.00	0.97±0.01	0.98±0.00	0.97±0.01	0.84±0.06	0.93±0.01	0.82±0.05	0.82±0.01
+ friction	0.99±0.00	0.97±0.01	0.99±0.00	0.97±0.01	0.81±0.07	0.94±0.01	0.80±0.06	0.80±0.04
Pendulum	0.94±0.01	0.90±0.02	0.93±0.00	0.91±0.01	0.75±0.04	0.78±0.02	0.69±0.03	0.60±0.06
+ friction	0.97±0.01	0.93±0.01	0.96±0.01	0.93±0.01	0.76±0.03	0.79±0.02	0.71±0.04	0.63±0.07
Double pendulum	0.90±0.01	0.90±0.02	0.88±0.01	0.89±0.02	0.63±0.03	0.70±0.06	0.56±0.03	0.40±0.09
+ friction	0.91±0.01	0.91±0.02	0.90±0.01	0.89±0.02	0.62±0.04	0.72±0.06	0.54±0.03	0.43±0.08
Two-body	1.00±0.00	0.99±0.01	0.99±0.00	0.98±0.01	0.97±0.01	0.97±0.01	0.94±0.01	0.87±0.07
+ friction	1.00±0.00	0.98±0.01	0.99±0.00	0.98±0.01	0.97±0.01	0.97±0.01	0.94±0.01	0.90±0.03
Mean	0.96±0.04	0.95±0.03	0.95±0.04	0.94±0.04	0.79±0.13	0.85±0.11	0.75±0.14	0.68±0.18

search among text candidates is easier; the same asymmetry appears in the λ=0 cycles (Table 17: text 0.83, vision 0.43), and it is consistent with the shared information of App. D.4 (5–13 bits against the $\log_2 450 \approx 8.8$ that identification needs). In both directions the reference multiplies retrieval several-fold and the two conditions do not overlap; the same holds for nonlinear and invertible maps (Table 21).

Table 19: Mapped text→vision r@1 per system with the fixed penalty (α=1) and with α selected for R^2 on an inner split (mean±std over seeds). The λ=0 floor holds on every system and tuning lowers it slightly; two-body is the highest λ=0 cell and the lowest auxiliary cell.

	aux		no-aux	
System	α=1	tuned	α=1	tuned
Mass–spring	0.85±0.03	0.85±0.03	0.03±0.01	0.02±0.01
+ friction	0.84±0.03	0.84±0.02	0.03±0.01	0.02±0.01
Pendulum	0.90±0.02	0.89±0.02	0.03±0.01	0.03±0.01
+ friction	0.90±0.03	0.89±0.02	0.04±0.01	0.02±0.01
Double pendulum	0.90±0.03	0.88±0.03	0.04±0.01	0.03±0.01
+ friction	0.91±0.03	0.89±0.04	0.04±0.01	0.02±0.01
Two-body	0.73±0.12	0.73±0.12	0.08±0.02	0.05±0.01
+ friction	0.73±0.11	0.73±0.12	0.08±0.01	0.06±0.02

Physical-symmetry test (full). We probe whether the induced correspondence encodes the *physical latent* or surface *appearance* by applying the frozen encoders and a base-fit ridge probe to two counterfactual twins of a held-out two-body system. The *color-swap* twin exchanges the discs' colors with physics held fixed, a true symmetry, so a physics-respecting representation should be *invariant* to it, and it is, in both conditions and modalities: latent cosine 0.99 (vision) / 1.00 (text), base-vs-twin energy-decode consistency 0.99/1.00, and cross-modal map consistency 0.99 ($v \to t$) / 1.00 ($t \to v$); neither encoder uses color. The *mass-swap* twin exchanges $m_1 \leftrightarrow m_2$, a genuine change of physics, so a faithful representation should be *equivariant*. Under the shared reference it is: decoded masses track the swap (mass-equivariance R^2 0.99 vision / 0.95 text) and the cross-modal map recovers the swapped masses from the other modality (R^2 0.98 $v \to t$, 0.97 $t \to v$ vs. ground truth). The purely generative (λ=0) model is markedly worse, mass-equivariance R^2 0.63/0.79 and cross-modal 0.76/0.77, and its latent barely moves under the swap (cosine 1.00), i.e. it does not represent the mass distinction the swap targets (Fig. 3). The induced map thus transports the *parameters* (equivariant to physics) and discards *appearance* (invariant to color; Fig. 7), and only the shared reference makes that transport cross-modal (Fig. 4).

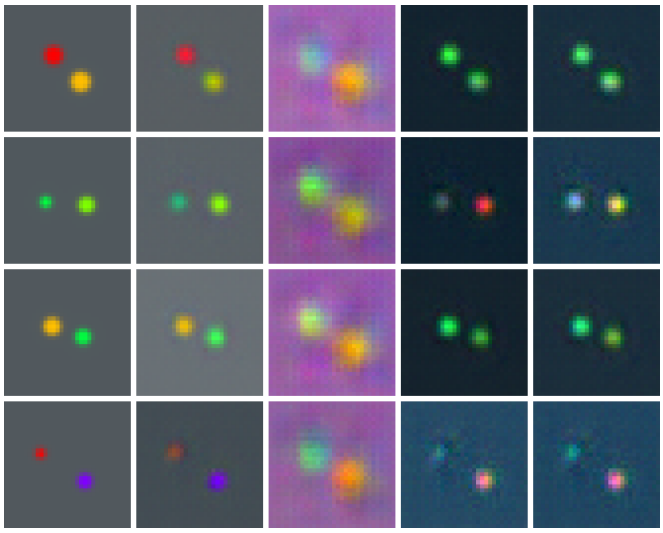

Figure 7: **Appearance invariance (color-swap).** The twin swaps the discs' colors with the physics unchanged. Rows: four twin trajectories. Columns, decoded by the frozen D_V: original; native latent $D_V(z_v)$; text latent $D_V(z_t)$; mapped text $D_V(W_{t\to v}z_t)$; round trip $D_V(W_{t\to v}W_{v\to t}z_v)$. The mapped scene keeps positions and sizes but not colors, and is unchanged between base and twin (consistency 0.99–1.00): the correspondence carries the physics and discards appearance.

E.2 Nonlinear Map Search

To rule out that the $\lambda{=}0$ failure is a limit of *mapper* capacity rather than of the representations, we fit an increasingly flexible battery of text→vision maps on a $70/30$ held-out split and report the best held-out r@1 over the battery: least-squares ridge; RBF kernel ridge swept over bandwidth ($\{0.1, 0.5, 1, 2\}/\mathrm{median}$) × ridge penalty ($\{0.1, 1, 10\}$); RFF kernel CCA (512 random-Fourier features, 32 canonical components); and deep MLPs swept over depth $\{2, 3, 4\}\times$ width $\{512, 1024\}\times$ weight-decay $\{0, 10^{-4}, 10^{-2}\}$. At $\lambda{>}0$ a plain ridge already recovers the correspondence and the nonlinear maps add little; at $\lambda{=}0$ the entire battery stays low: the best nonlinear map reaches r@1 0.10, about $9\times$ below the $\lambda{>}0$ linear map (Table 20).

Table 20: Nonlinear map search: mapped text→vision r@1 (mean over 8 systems, chance ≈ 0.002). Even the most flexible maps fail to recover the text→vision correspondence from the $\lambda{=}0$ encoders, while at $\lambda{>}0$ a plain linear map already suffices.

map family	aux ($\lambda{>}0$)	no-aux ($\lambda{=}0$)
ridge (linear)	$0.84{\pm}0.05$	$0.04{\pm}0.01$
kernel ridge (RBF)	$0.85{\pm}0.04$	$0.05{\pm}0.01$
kernel CCA (RFF)	$0.88{\pm}0.02$	$0.10{\pm}0.01$
deep MLP	$0.85{\pm}0.05$	$0.07{\pm}0.01$
best over battery	$0.90{\pm}0.02$	$0.10{\pm}0.01$

Both directions, including an invertible map. The battery above is text→vision. Table 21 refits ridge, a $2{\times}512$ MLP and a normalizing flow (four affine coupling layers on top of a linear layer initialized at the ridge solution; exactly invertible, $\approx$ 1.6M parameters) in both directions, on a $1050/150/300$ train/validation/test split. For every map family the reference condition is several-fold above $\lambda{=}0$, and the flow does no better than ridge: text→vision stays at the floor (0.04–0.06), and vision→text, the easier direction (App. E.1), reaches 0.30–0.36 against 0.92–0.95. The two $\lambda{=}0$ spaces are therefore not reparameterizations of one another.

Table 21: Linear, nonlinear and invertible maps in both directions: held-out r@1 and R^2 of the fitted map (mean-pooled latents; mean over the eight systems, then mean±std over five seeds); best map per column and metric in bold. Retrieval is among 300 test rows (chance $1/300$), a smaller gallery than in Table 16, which is why the ridge values are slightly higher here.

		aux (λ>0)		no-aux (λ=0)	
metric	map	text→vision	vision→text	text→vision	vision→text
r@1	ridge	0.87±0.04	0.94±0.03	**0.06±0.01**	0.30±0.02
	MLP (2×512)	**0.89±0.04**	**0.95±0.02**	**0.06±0.01**	0.34±0.03
	flow (4 couplings)	0.86±0.04	0.92±0.03	0.04±0.01	**0.36±0.02**
R^2	ridge	0.76±0.02	0.86±0.01	**0.13±0.03**	0.39±0.04
	MLP (2×512)	**0.80±0.02**	**0.89±0.00**	0.07±0.07	**0.50±0.02**
	flow (4 couplings)	0.74±0.02	0.80±0.01	0.03±0.03	0.45±0.02

E.3 Post-hoc alignment baseline (sample complexity)

We fit Procrustes, CCA, and ridge maps between the *frozen* encoders on k paired anchors and sweep k up to the full pool (1050), for both the no-auxiliary and auxiliary models (Table 22). No-auxiliary alignment does not rise at any budget, so the failure is not a shortage of paired supervision; auxiliary alignment climbs with k, and the free-linear ridge map far exceeds the orthogonal Procrustes map, which independently indicates a general linear correspondence rather than a rotation.

E.4 Neighbourhood size and latent geometry

We report three geometry diagnostics that accompany the alignment gap. The mutual-kNN advantage holds across neighbourhood size k, not just the k=10 used in the main text (Table 23); under the shared reference the two latents' intrinsic dimension (participation ratio) converges to a common value while at λ=0 they differ by $\sim 2.8\times$ (Table 24); and the fitted map's non-orthogonality persists after whitening (Table 25).

Table 22: Post-hoc alignment from k paired anchors between the *frozen* encoders: cross-modal recall-at-1 by Procrustes, CCA and ridge (α=10), mean over eight systems, chance ≈ 0.002. The final row switches metric to mutual-kNN@10 at the full anchor budget, so its "10" counts neighbours and its "1050" anchors.

	no-aux (λ=0)			aux (λ>0)		
anchors k	Procr.	CCA	ridge	Procr.	CCA	ridge
10	0.009±0.004	0.009±0.004	0.010±0.005	0.077±0.037	0.099±0.041	0.125±0.057
50	0.016±0.008	0.012±0.004	0.020±0.008	0.261±0.060	0.087±0.025	0.448±0.068
200	0.017±0.011	0.018±0.006	0.030±0.012	0.351±0.052	0.194±0.052	0.700±0.084
1050	0.014±0.009	0.025±0.009	0.043±0.025	0.380±0.057	0.304±0.083	**0.832±0.082**
mkNN@10 (all 1050 anchors)	0.079±0.029	0.080±0.030	0.100±0.029	0.483±0.096	0.286±0.053	**0.642±0.085**

Table 23: Mutual-kNN overlap vs neighbourhood size k (mean over 8 systems). This is a standalone k-sweep whose absolute values are *not* matched to the seed-averaged mutual-kNN@10 quoted in the main text (Tables 28, 10: 0.41 aux / 0.047 no-aux); it is reported only to show that the aux ≫ no-aux ordering, and the monotone growth with k, hold at every neighbourhood size.

k	1	5	10	25	50	100
aux	0.287±0.054	0.431±0.079	0.477±0.089	0.526±0.099	0.556±0.101	0.602±0.094
no-aux	0.025±0.012	0.053±0.024	0.074±0.029	0.122±0.036	0.189±0.036	0.298±0.033

Table 24: Intrinsic dimensionality: participation ratio $(\sum \sigma_i)^2 / \sum \sigma_i^2$ and #dims for 90%/99% variance, with vs without the target (mean over 8 systems).

modality	cond.	part. ratio	dims@90%	dims@99%
vision	aux	6.58±3.00	12.9±5.8	61.4±18.7
vision	no-aux	4.50±1.11	9.0±4.4	53.1±27.3
text	aux	9.21±3.78	12.3±5.3	47.2±11.7
text	no-aux	12.57±2.87	32.3±7.9	121.0±20.6

Table 25: Orthogonality deviation of the fitted text→vision map on raw, z-scored (per-dimension scale removed), and ZCA-whitened (anisotropy *and* cross-dimension correlation removed) latents (aux; mean over 8 systems). Whitening lowers it but leaves the map far from orthogonal (0 = scaled rotation).

latents	raw	z-scored	ZCA-whitened
ortho_dev	6.70±2.16	3.80±0.45	1.87±0.06

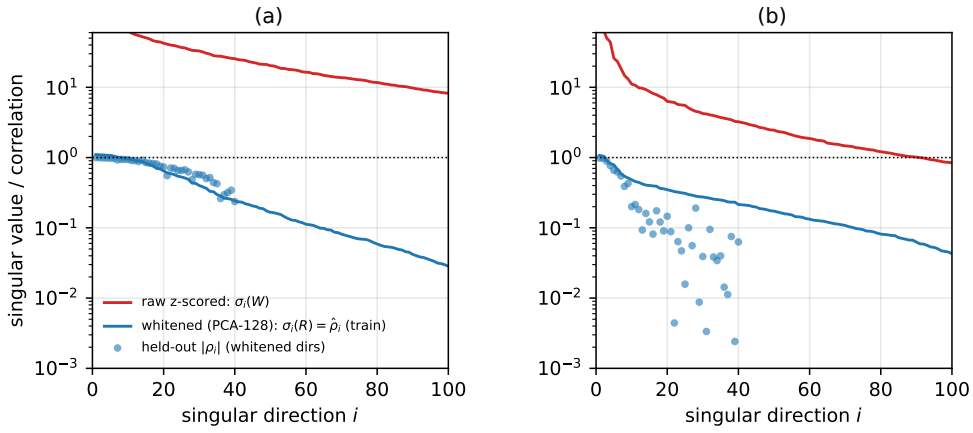

Figure 8: Singular values of the fitted text→vision map on the pendulum pair, aux (a) and λ=0 (b), in order; a rotation would sit on the dotted line at 1. Raw z-scored coordinates in red; after PCA-128 and shrinkage whitening of both spaces (blue) the spectrum is the canonical-correlation spectrum, with held-out correlations as dots. The aux pair has five directions at $\rho^2 > 0.9$ against two at λ=0.

F What the Target Must Carry

This appendix gives the full results for §6.3.

F.1 Auxiliary-target ablations

We probe every component of the auxiliary objective: the phase head and reconstruction normalization (Table 26), the fixed-γ balancer sweep (Table 31), a minimal single symmetric target, either system identity or scalar energy alone (Table 29), the partial-overlap spectrum including disjoint and identical targets (Table 30), and the dimension of the H-subspace each condition installs (Table 32). Together these show that the effect survives with a single scalar target and without the dense phase head, and increases as target information about h increases.

Table 26: Auxiliary-configuration ablations (mean over the eight base systems). Removing the per-frame phase head leaves alignment far above the no-auxiliary floor; normalizing the reconstruction term changes it negligibly. The "no-aux (floor)" row here ($0.17/0.04/0.03$) sits slightly below the seed-averaged floor ($0.21/0.047/0.042$; Table 10); the gap to the aux rows is the point.

ablation	variant	CKA	kNN@10	r@1
–	no-aux (floor)	0.17	0.04	0.03
phase head	aux, no phase head	0.600	0.207	0.542
	aux, + phase head	**0.639**	**0.220**	**0.677**
recon norm.	unnormalized reconstruction	0.600	0.207	0.542
	standard reconstruction	0.581	0.173	0.465

Table 27: Held-out generative quality with and without the shared reference (Latte-S, mean±std over seeds and the eight base systems). The reference costs some generative quality in both modalities, the Pareto trade-off of Eq. (4), so the alignment it installs does not come from better generation.

	vision ϵ-MSE	vision PSNR (dB)	text NLL	text perplexity
aux (λ>0)	0.0027±0.0005	28.8±0.4	0.568±0.019	1.77±0.04
no-aux (λ=0)	0.0013±0.0007	30.7±1.9	0.443±0.001	1.56±0.00

Table 28: Target variants, mean over eight systems; best per column in bold. Row labels state whether each target is shared and whether it carries h; *asymmetric* means vision regresses energy only and text parameters only. View-matching couples the encoders directly and uses no h (App. B); the full partial-overlap sweep is in Table 30.

target	CKA	norm. CKA	mutual-kNN	mapped r@1
none (λ=0)	0.21±0.07	0.23±0.08	0.047±0.019	0.042±0.023
shuffled target (shared, $\perp h$)	0.19±0.07	0.27±0.10	0.044±0.020	0.038±0.020
VICReg (view-match, couples encoders)	0.54±0.18	0.68±0.23	0.061±0.031	0.354±0.143
Barlow Twins (view-match, couples encoders)	0.68±0.11	**0.78±0.13**	0.052±0.016	0.050±0.020
rand. projection (shared, info-preserving)	0.48±0.15	0.58±0.18	0.154±0.084	0.378±0.158
asymmetric (not shared, each informative about h)	0.41±0.10	0.52±0.13	0.146±0.061	0.281±0.135
identical (shared, all of h)	**0.72±0.09**	0.77±0.10	**0.408±0.076**	**0.844±0.093**

F.2 Alignment is not an artifact of the auxiliary head's capacity

The auxiliary head that shapes each encoder is, by default, an attention-pool followed by a two-layer MLP. One might worry that the head, rather than the shared reference, produces the alignment through its own capacity. We rule this out by reducing the head to the minimal form the theory assumes: a purely affine map $\rho(e) = Ce + c$, i.e. a mean-pool followed by a single linear layer, with no attention and no nonlinearity. If alignment were a decoder artifact, this collapse should remove it. Averaged over the eight systems, the affine head instead attains raw CKA 0.735 ± 0.034

Table 29: Minimal single symmetric target (mean over 8 systems); best per column in bold. The single-target rows are trained without the per-frame phase head; the last row is the default full target, which includes it. A lone *scalar* target (energy) lifts alignment well above the no-auxiliary floor, whereas a lone *categorical* target (system identity) does not.

shared target	CKA	kNN@10	r@1
no-aux (λ=0)	0.21±0.07	0.047±0.019	0.04±0.02
system identity only	0.14±0.04	0.05±0.02	0.02±0.01
energy (scalar) only	0.48±0.07	0.11±0.04	0.20±0.07
params+energy+system+phase (default)	**0.72±0.09**	**0.41±0.08**	**0.84±0.09**

Table 30: Partial target overlap (mean over 8 systems); best per column in bold. Vision and text predict group sets with varying overlap. The identity-level metrics stay flat across the partial rows, which share only a scalar, and rise when the full target is shared. The *disjoint* row assigns vision→params, text→energy, the opposite of Table 28's asymmetric row.

overlap (vision / text)	CKA	kNN@10	r@1
disjoint (params / energy)	0.45±0.09	0.152±0.076	0.39±0.15
share energy	0.55±0.10	0.159±0.075	0.42±0.15
share energy+system	0.60±0.07	0.154±0.065	0.40±0.14
full (matched)	**0.72±0.09**	**0.41±0.08**	**0.84±0.09**
no-aux	0.21±0.07	0.047±0.019	0.04±0.02

Table 31: Fixed-γ ablation; best per column in bold. γ is the fraction of the loss on reconstruction, so the auxiliary target gets weight $1-\gamma$. Alignment stays far above the no-auxiliary floor (0.21/0.047/0.042) at every setting, including γ=0.9 (only 10% weight on the target), so it does not require a specific recon / aux balance; the adaptive balancer sits near the optimum.

balancer (γ, aux frac $1-\gamma$)	CKA	kNN@10	r@1
adaptive (baseline)	**0.72±0.09**	**0.41±0.08**	0.84±0.09
γ=0.1 (0.9)	0.70±0.08	0.28±0.04	0.67±0.14
γ=0.3 (0.7)	0.69±0.07	0.38±0.07	**0.85±0.07**
γ=0.5 (0.5)	0.65±0.10	0.36±0.08	0.78±0.12
γ=0.7 (0.3)	0.70±0.08	0.33±0.05	0.71±0.14
γ=0.9 (0.1)	0.64±0.09	0.34±0.08	0.69±0.11

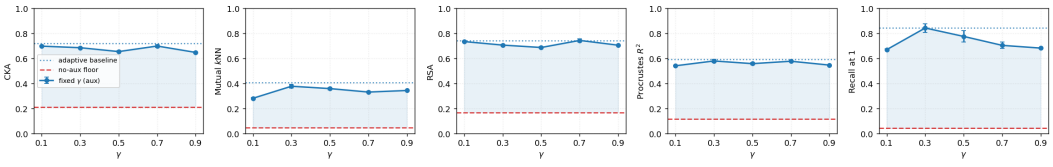

Figure 9: Fixed-γ balancer sweep across all five alignment measures (mean over 8 systems; error bars are seed std). Alignment is flat across $\gamma \in [0.1, 0.9]$, staying near the adaptive baseline (dotted) and far above the no-auxiliary floor (dashed) on every measure, the dissociation does not depend on the reconstruction/auxiliary balance.

Table 32: Near-unit canonical count against the shared target. Shared dims: within-system-varying components of h both heads regress (full target: 5–12 per system). Near-unit count: held-out canonical correlations above the stated threshold, an estimate of k, not k itself. The first row is the frozen pretrained pairs, which shares no training world (App. G.6).

	shared	near-unit count	
target condition	dims	$\rho^2{>}0.8$	$\rho^2{>}0.9$
pretrained, zero-shot	0	0.0	0.0
none (λ=0)	0	2.7±1.3	1.9±0.8
disjoint	0	3.5±1.4	2.0±1.0
asymmetric	0	4.0±1.3	2.0±0.8
share energy	1	3.9±1.8	2.5±1.1
share energy + label	1	4.2±1.7	2.7±1.2
full (all of h)	5–12	9.3±1.9	5.6±1.5

(Table 33), on par with, and slightly above, the attention-pool+MLP head (0.692 ± 0.015) and $3.6\times$ the no-auxiliary floor (0.206 ± 0.029). The aligning geometry is therefore shaped in the *encoder* by the shared reference and is already expressed under a linear read-out, and a larger head does not increase it. The effect is stable across seeds (per-system std 0.027–0.069) and holds from near-integrable (two-body, 0.840) to chaotic (double pendulum, 0.633) dynamics (Table 34).

Table 33: Auxiliary-head capacity control. Cross-modal alignment (raw linear CKA, mean ± std over seeds) is unchanged when the auxiliary head is reduced from an attention-pool+MLP to a single affine map $\rho(e) = Ce + c$, and both are far above the no-auxiliary floor, showing that the head capacity does not induce alignment.

Auxiliary head	CKA
Affine $\rho(e) = Ce + c$ (mean-pool + linear)	0.735 ± 0.034
Attention-pool + MLP (default)	0.692 ± 0.015
None (λ=0)	0.206 ± 0.029

Table 34: Per-system cross-modal alignment (raw linear CKA) with the affine auxiliary head $\rho(e) = Ce + c$ (mean ± std over seeds).

System	CKA (mean ± std)
Mass–spring	0.726 ± 0.069
Mass–spring (friction)	0.739 ± 0.068
Pendulum	0.722 ± 0.029
Pendulum (friction)	0.742 ± 0.027
Double pendulum	0.633 ± 0.035
Double pendulum (fric.)	0.648 ± 0.036
Two-body	0.840 ± 0.062
Two-body (friction)	0.832 ± 0.060
Overall	$\mathbf{0.735 \pm 0.034}$

G Robustness and Scale

This appendix gives the robustness and scaling results behind §6.4.

G.1 Read-out compression

Figure 10 and Table 35 report the full K-query bottleneck sweep across all five alignment measures; the auxiliary model exceeds the no-auxiliary model at every width on every measure.

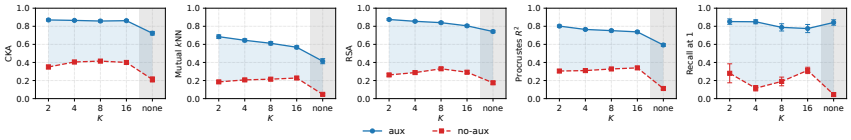

Figure 10: **Read-out compression.** Cross-modal alignment under a K-query bottleneck ($K \in \{2, 4, 8, 16\}$) or mean-pool ("none"), across five measures, mean over eight systems; aux (λ>0) in blue, λ=0 in red.

Table 35: **Read-out compression.** Cross-modal alignment under a K-query bottleneck ($K \in \{2, 4, 8, 16\}$) or mean pool ("none"), mean over eight systems.

	CKA		mutual-kNN		RSA		Procrustes R^2		Recall at 1	
Read-out	aux	no-aux	aux	no-aux	aux	no-aux	aux	no-aux	aux	no-aux
K=2	0.88±0.05	0.35±0.06	0.69±0.08	0.18±0.04	0.88±0.06	0.24±0.04	0.80±0.06	0.30±0.07	0.87±0.06	0.28±0.13
K=4	0.87±0.06	0.41±0.08	0.65±0.08	0.22±0.06	0.86±0.06	0.30±0.09	0.77±0.04	0.32±0.07	0.85±0.06	0.12±0.06
K=8	0.86±0.07	0.42±0.07	0.61±0.08	0.22±0.05	0.84±0.08	0.33±0.07	0.75±0.06	0.33±0.06	0.80±0.11	0.19±0.08
K=16	0.87±0.05	0.40±0.06	0.59±0.06	0.22±0.05	0.84±0.07	0.29±0.07	0.76±0.05	0.34±0.06	0.80±0.12	0.31±0.12
none	0.72±0.09	0.21±0.07	0.41±0.08	0.047±0.019	0.69±0.01	0.15±0.03	0.55±0.01	0.10±0.01	0.84±0.09	0.042±0.023

G.2 Text read-out and caption length

The text latent is a pooled summary of GPT-2's hidden states. We re-pool the same encoders three task-agnostic ways and re-align against the fixed vision latent (Table 36); a stronger aggregator does not recover no-auxiliary alignment.

Table 36: Text read-out sensitivity: the same encoders re-pooled three ways, re-aligned against the fixed vision latent (mean over 8 systems; seed-averaged values in Table 10).

	aux			no-aux		
read-out	CKA	kNN	r@1	CKA	kNN	r@1
mean	0.73±0.09	0.477±0.089	0.78±0.09	0.23±0.07	0.074±0.029	0.04±0.02
last-token	0.61±0.11	0.358±0.085	0.60±0.15	0.13±0.10	0.046±0.016	0.04±0.02
max	0.75±0.07	0.448±0.066	0.59±0.11	0.21±0.05	0.062±0.022	0.02±0.01

Table 37: Caption length. Doubling the caption budget to 512 tokens (denser interleaved phase-space samples) leaves the dissociation unchanged: the auxiliary model stays high and the no-auxiliary model stays at the floor, matching the 256-token default within seed noise (Latte-S, mean over 8 systems).

caption length	cond.	CKA	mutual-kNN	r@1
256 (default)	aux	0.72±0.09	0.41±0.08	0.84±0.09
512	aux	0.71±0.09	0.40±0.07	0.81±0.10
256 (default)	no-aux	0.21±0.07	0.047±0.019	0.04±0.02
512	no-aux	0.23±0.07	0.048±0.021	0.04±0.02

G.3 Vision-architecture ablation

To test whether the dissociation is specific to the Latte diffusion backbone, we swap the vision encoder for an alternative latent-video model at *matched* scale (d=384, width-matched to GPT-2, the same as Latte-S) VideoFlow-S (Kumar et al., 2020) holding the text tower, the auxiliary target (params+energy+system, no per-frame phase), and every optimization detail fixed. It is evaluated

in the $\lambda>0$ and $\lambda=0$ conditions against the shared text encoder, with the identical alignment pass on the eight base systems (Table 38). The aux-vs-no-aux dissociation replicates: alignment is high under the shared reference and at the floor without it, so the effect does not depend on the Latte backbone.

Table 38: Vision-architecture ablation: cross-modal alignment for an alternative latent-video backbone at matched Latte-S scale (d=384), aux ($\lambda>0$) vs no-aux ($\lambda=0$), mean over the eight base systems. Swapping Latte for a different latent-video model reproduces the dissociation, alignment is high under the shared reference and stays at the $\lambda = 0$ level without it.

	aux ($\lambda>0$)			no-aux ($\lambda=0$)		
Vision backbone	CKA	kNN	r@1	CKA	kNN	r@1
Latte-S/4	0.72±0.09	0.41±0.08	0.84±0.09	0.21±0.07	0.047±0.019	0.04±0.02
VideoFlow-S	0.77±0.08	0.46±0.03	0.86±0.09	0.07±0.03	0.04±0.01	0.01±0.01

G.4 Text-objective ablation

To test whether alignment depends on the *form* of the text objective, we retrain the text encoder under three objectives (autoregressive GPT-2, score-entropy diffusion (SEDD) (Lou et al., 2024), and masked diffusion (MDLM) (Sahoo et al., 2024)) each with and without the auxiliary target, against the same diffusion vision encoder (Table 39). Every objective aligns under the shared target and none aligns without it. Matching the vision model's diffusion form (SEDD/MDLM) lowers the $\lambda=0$ floor, so objective similarity does not produce the correspondence.

Table 39: Text-objective ablation. Three text objectives (autoregressive GPT-2, score-entropy diffusion SEDD, masked diffusion MDLM), each × aux/no-aux, against the same diffusion vision (mean over eight systems). Every objective aligns under the shared target and none aligns without it; matching the vision model's diffusion form (SEDD/MDLM) lowers the $\lambda=0$ floor.

	CKA		kNN		r@1	
text objective	aux	no	aux	no	aux	no
GPT-2 (autoregressive)	0.72±0.09	0.21±0.07	0.41±0.08	0.047±0.019	0.84±0.09	0.042±0.023
SEDD (score-entropy diff.)	0.73±0.11	0.09±0.03	0.37±0.04	0.03±0.01	0.81±0.09	0.03±0.01
MDLM (masked diffusion)	0.73±0.10	0.13±0.06	0.38±0.05	0.03±0.01	0.81±0.08	0.03±0.01

G.5 Scale and data scaling

Table 40 reports alignment across backbone scale (Latte-S/B/L), and Table 41 the effect of scaling the training set to 80k trajectories at fixed capacity; the no-auxiliary control stays at the floor throughout. Table 42 reports alignment across training on 2, 4, and 8 distinct systems at a *matched* total-trajectory budget ($\sim$ 48k: 24k/12k/6k per system) and fixed epochs, so the number of systems is the only variable; all three models are evaluated on the same common held-out systems. Alignment is flat across the sweep, so diversity across system families does not induce alignment either, and the diversity axis on its own cannot explain the misalignment.

G.6 Zero-shot pretrained pairs

The sweeps above vary capacity and data within our training setup; this one removes the setup entirely. We freeze four pairs of models pretrained on natural data (Table 44): Latte-B/2 or Latte-L/2 (Ma et al., 2025) on FaceForensics (Rössler et al., 2018) for video, each paired with a text model of exactly matched width, depth and head count, GPT-2 (Radford et al., 2019) on WebText or SEDD (Lou et al., 2024) on OpenWebText (Gokaslan & Cohen, 2019), and run the same alignment pass on our eight systems. Within a pair the two models share no training data, no objective and no tokenizer, and none has seen a rendered trajectory or a caption from our generator. Read-outs are mean pools

of the final hidden states, captions are retokenized by each text model's vocabulary, and maps and probes use the same $70/30$ split. A frozen pair has no seeds, so spreads are over the eight systems and the ceiling-normalized CKA of App. C is undefined.

The pretrained encoders represent the physics, and the two latents share almost nothing (Table 43). On the video side a held-out probe recovers the Hamiltonian parameters as well as our λ=0 encoder does; on the text side the pretrained encoders are weaker (0.53–0.65 against 0.85), with the gap concentrated on the double pendulum, whose captions a model not trained on them reads least well. Yet every map-free measure sits at or below the λ=0 floor, and no canonical direction reaches $\rho^2 = 0.9$ in any system for any pair, against 1.9 at λ=0 and 5.6 under the full target. Retrieval alone runs the other way, because it is computed after a map fit on 70% of the pairs: it tests whether a linear relation between the spaces can be recovered from paired examples, a weaker question than whether they agree, and the wider pretrained latents (d=768 and 1024, against 384) give that fit more to work with. It remains about seven times below the full-target condition. Our λ=0 encoders keep a small H-subspace, while the pretrained pairs have none ($k = 0$ in the sense of Definition 1).

Table 40: Alignment across backbone scale (mean±std over seeds, S: 5, B: 2, L: 1). No-auxiliary CKA stays at the floor across a $2.7\times$ width range while the auxiliary model holds.

		CKA		mutual-kNN		mapped r@1	
Backbone	d	aux	no-aux	aux	no-aux	aux	no-aux
Latte-S	384	0.72±0.09	0.21±0.07	0.41	0.05	0.84	0.04
Latte-B	768	0.72±0.09	0.23±0.07	0.33	0.05	0.65	0.05
Latte-L	1024	0.70	0.08	0.39	0.02	0.76	0.05

Table 41: Data scaling at fixed capacity (Latte-S, mean over eight systems). Scaling the training set to 80k trajectories leaves the no-auxiliary model at the floor; the gap does not narrow with data, while the auxiliary model stays high.

data	cond.	CKA	mutual-kNN	r@1
48k	no-aux	0.21±0.07	0.047±0.019	0.04±0.02
80k	no-aux	0.17±0.08	0.032±0.019	0.05±0.03
48k	aux	0.72±0.09	0.41±0.08	0.84±0.09
80k	aux	0.74±0.10	0.378±0.090	0.79±0.12

Table 42: No-auxiliary alignment vs. training breadth (matched $\sim$48k total trajectories, 30 epochs; mean over the common held-out systems mass–spring and two-body). Alignment is flat from two to eight training datasets: added diversity alone does not induce alignment.

training breadth	CKA	mutual-kNN	mapped r@1
2 systems (24k/sys)	0.243±0.034	0.059±0.004	0.060±0.016
4 systems (12k/sys)	0.221±0.053	0.048±0.002	0.070±0.025
8 systems (6k/sys)	0.244±0.046	0.057±0.012	0.050±0.012

Table 43: Alignment between independently pretrained encoders, mean±std over eight systems; ρ_1 and decode are means over systems. Chance r@1 $= 0.0022$, mutual-kNN $= 0.0067$. ρ_1 is the top held-out canonical correlation ($\rho_1^2 \leq 0.83$ for every pair); no pair has a direction with $\rho^2 > 0.9$ in any system. *Decode* is the held-out ridge R^2 of the Hamiltonian parameters from each latent: the video encoders recover them at $R^2 \approx 0.85$ and the text encoders at 0.53–0.65, so the low alignment is not an absence of information.

pair	CKA	mutual-kNN	RSA	Procr. R^2	r@1 $t\to v$	ρ_1	decode v/t
B / GPT-2	0.057±.018	0.018±.004	0.035±.028	0.048±.008	0.121±.023	0.88	0.85/0.65
B / SEDD	0.075±.015	0.027±.005	0.044±.027	0.062±.006	0.114±.032	0.90	0.85/0.53
L / GPT-2	0.058±.018	0.017±.003	0.039±.031	0.053±.009	0.114±.024	0.85	0.84/0.65
L / SEDD	0.122±.030	0.044±.019	0.083±.036	0.098±.013	0.099±.018	0.91	0.84/0.65

Table 44: Zero-shot pretrained pairs: video and text encoders trained independently on unrelated corpora, never on our data, with matched width, depth and heads. The video path also carries a Stable Diffusion VAE trained on LAION (Schuhmann et al., 2022).

pair	video encoder	pretrained on	text encoder	pretrained on
B / GPT-2	Latte-B/2	FaceForensics	GPT-2 small	WebText
B / SEDD	Latte-B/2	FaceForensics	SEDD small	OpenWebText
L / GPT-2	Latte-L/2	FaceForensics	GPT-2 medium	WebText
L / SEDD	Latte-L/2	FaceForensics	SEDD medium	OpenWebText
width/depth/heads: B pairs 768/12/12, L pairs 1024/24/16, both modalities				

H HELD-OUT SYSTEMS

We run two out-of-distribution experiments that probe complementary axes of generalization. The **three-body** system is OOD in the *number of interacting bodies* and is chaotic; the **charged two-body** (Coulomb) system is OOD in the *force law* itself, a sign-varying interaction (Hamiltonian in Table 3) that, unlike gravity, can either attract or repel. Encoders are frozen and never trained on either system. Full per-system metrics are in Table 47.

The gap survives on Coulomb; on three-body it is weak in map-free CKA but clear in retrieval. On three-body, raw CKA separates only modestly (0.59 aux vs. 0.56 no-aux), while the auxiliary model keeps a clear *retrieval* advantage (0.26 vs. 0.05). These measures dissociate because they probe different structure: CKA is map-free and reads coarse second-order covariance both conditions capture, whereas retrieval relies on the fitted map, which only the auxiliary code populates with a fine per-trajectory identity. On the charged system, which shares the central-force two-body geometry the encoders were trained on, the coarse component is target-dependent too, so both measures separate (aux CKA 0.41, r@1 0.49 vs. no-aux 0.22/0.02).

The learned map is moderate-rank across regimes. We measure the effective rank $(\sum \sigma_i)^2 / \sum \sigma_i^2$ of the fitted text→vision map $W_{t\to v}$ (Table 45). The auxiliary map is not degenerately low-rank (≈ 61–74 of 384) and stays above the no-auxiliary map (≈ 53–60) in every regime; the rank gap the shared reference buys is modest in distribution (+5) and widens on the two held-out systems (+14 three-body, +20 Coulomb), consistent across seeds. On three-body the round-trip itself does not dissociate (both conditions return on-manifold, aux 0.84, no-aux 0.92 on the text cycle), so the fitted maps are well-formed there and the gap lies in the cross-modal correspondence.

Degradation & floor-pinned control. Across both systems the auxiliary alignment degrades from the in-distribution level (aux CKA 0.72 → 0.41 on Coulomb, → 0.59 on three-body) while the no-auxiliary control never leaves the floor on the identity-level metrics (r@1 ≈ 0.02–0.05, mutual-kNN ≈ 0.03–0.05), even where the coarse CKA gap narrows. On the charged system the gap even survives read-out compression: the K=8 bottleneck retains an intermediate separation (aux CKA 0.31, r@1 0.10; no-aux 0.18/0.03).

Qualitative transfer. The stitching and round-trip visualizations corroborate the metrics. On three-body the frozen video decoder renders coherent scenes from its *own* latent (Fig. 11), yet a caption stitched through the map $D_V(W_{t\to v}g(y))$ only partially recovers the scene (Fig. 12), so the shortfall lies in the cross-modal correspondence rather than in generation. On the charged system the round-trip map $W_{t\to v}W_{v\to t}$ stays near-identity with the shared reference (Fig. 13) but becomes asymmetric without it (Fig. 14), the same failure mode seen in distribution.

Table 45: Effective rank $(\sum \sigma_i)^2/\sum \sigma_i^2$ of the fitted text→vision map $W_{t\to v}$ (out of d=384), mean±std over seeds. The auxiliary map is moderate-rank and stays above the no-auxiliary floor in every regime; the rank gap the shared reference buys widens out of distribution.

regime	aux (λ>0)	no-aux (λ=0)
in-distribution (two-body)	61±4	56±3
three-body (chaotic, OOD)	74±4	60±4
charged two-body (Coulomb, OOD)	73±3	53±5

Table 46: Asymmetric-target transfer to the held-out systems: the identical target, the *asymmetric* target (vision regresses energy only, text parameters only), and the no-auxiliary control. The asymmetric target keeps a retrieval edge over the floor but stays below the identical target, the thinner, low-rank correspondence a scalar anchor predicts (App. I.4). Asymmetric rows: 5 seeds; identical and no-aux rows: the frictionless rows of Table 47.

OOD system	target	CKA	mutual-kNN	r@1
Three-body	identical	0.58±0.04	0.11±0.02	0.26±0.05
	asymmetric	0.50±0.08	0.07±0.01	0.08±0.02
	no-aux (λ=0)	0.55±0.04	0.05±0.00	0.05±0.01
Coulomb	identical	0.41±0.04	0.19±0.03	0.49±0.11
	asymmetric	0.27±0.09	0.06±0.01	0.09±0.03
	no-aux (λ=0)	0.22±0.06	0.03±0.00	0.02±0.01

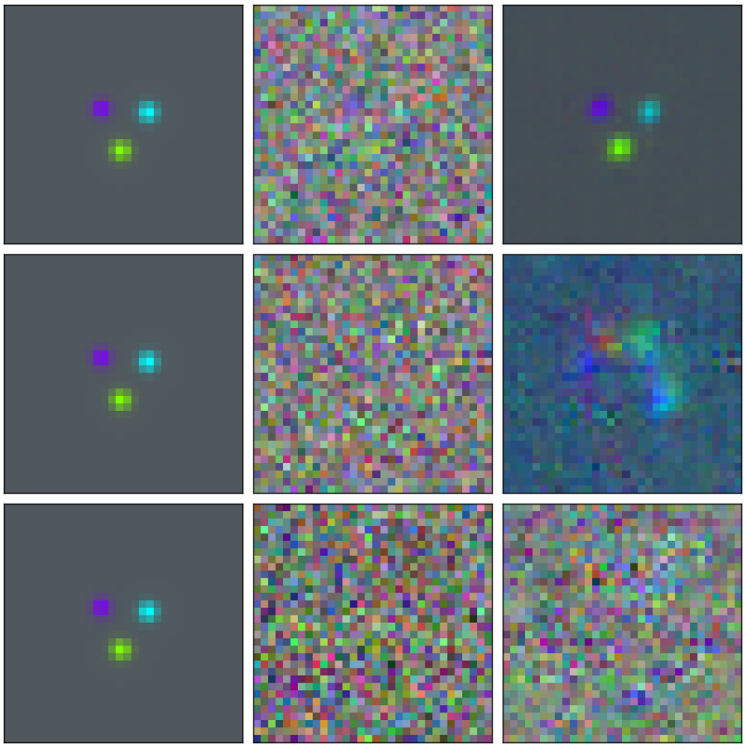

Figure 11: **Generative reconstruction on three-body.** Columns: original middle frame; noised input x_τ; denoised prediction $\hat{x}_0$ from the video model's own latent. Rows: diffusion step $\tau = 149,\ 499,\ 949$ (of 1000). At low noise the model renders a coherent three-body scene from the native latent, so the degraded *cross-modal* stitch (Fig. 12) reflects a missing text→vision correspondence; the decoder itself generates the scene.

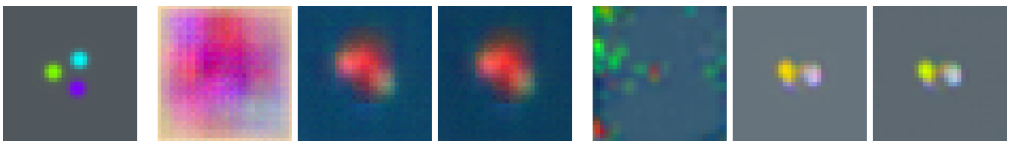

Figure 12: **Cross-modal stitching, three-body (zero-shot).** Ground truth (1), then the no-auxiliary (2–4) and auxiliary (5–7) models decoded by the frozen video decoder D_V three ways: direct text latent $D_V(z_t)$ (2, 5), mapped text $D_V(W_{t\to v} z_t)$ (3, 6), and the round-trip $D_V(W_{t\to v} W_{v\to t} z_v)$ (4, 7). Even with the shared reference the scene is only partially recovered.

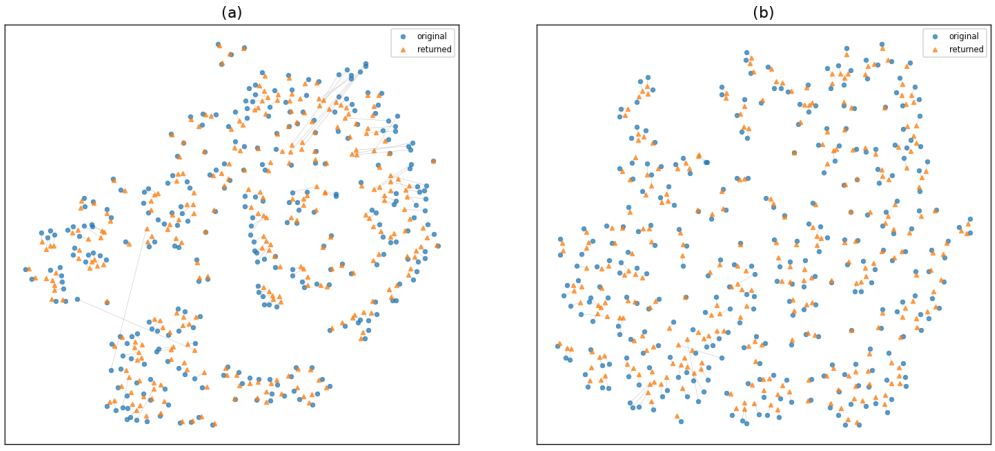

Figure 13: **Round-trip map, charged two-body (Coulomb), auxiliary ($\lambda>0$).** t-SNE of original (circles) and returned (triangles) latents, lines joining each pair: (a) text cycle $t\to v\to t$ (cycle r@1 0.92, R^2 0.84); (b) vision cycle $v\to t\to v$ (r@1 0.91, R^2 0.78). The cycle stays near-identity on the held-out force law, so the induced correspondence transfers zero-shot to the novel Coulomb interaction.

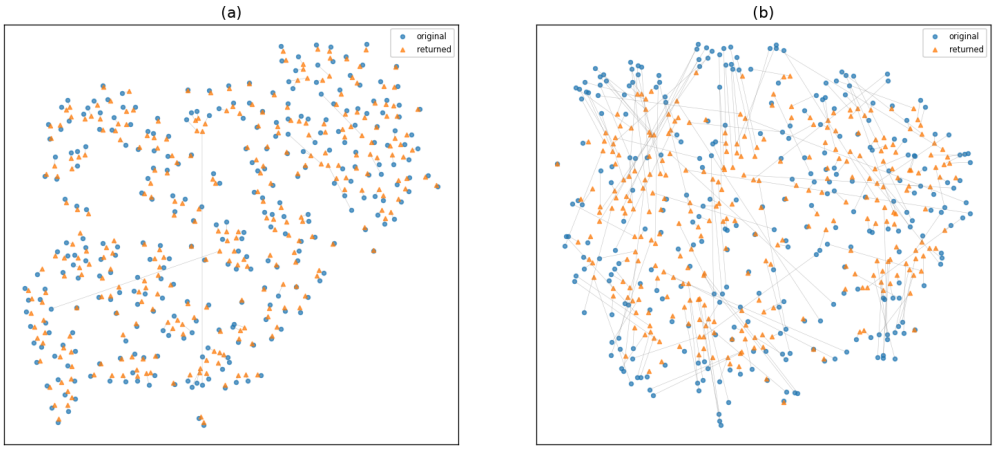

Figure 14: **Round-trip map, charged two-body (Coulomb), no-auxiliary ($\lambda=0$).** Panels as in Fig. 13: (a) text cycle (r@1 0.98, R^2 0.79); (b) vision cycle (r@1 0.68, R^2 0.65). The vision cycle degrades away from identity while the text cycle stays near-identity, the same asymmetric failure as in distribution, and the direction whose retrieval collapses without the shared reference.

H.1 PER-SYSTEM METRICS

Table 47 gives the full per-system zero-shot out-of-distribution alignment across all five measures (the summary appears in Table 12).

Table 47: Zero-shot out-of-distribution alignment, per held-out system. The affine-invariant SVCCA does not dissociate (on three-body it slightly favors no-aux), which is why we exclude it from the convergence claim; raw CKA on three-body is likewise close (aux $\approx$ no-aux), while the identity-level kNN and r@1 retain the gap.

	CKA		mutual-kNN		RSA		SVCCA		r@1	
System	aux	no-aux	aux	no-aux	aux	no-aux	aux	no-aux	aux	no-aux
Three-body	0.58±0.04	0.55±0.04	0.11±0.02	0.05±0.00	0.50±0.05	0.45±0.03	0.38±0.02	0.41±0.02	0.26±0.05	0.05±0.01
+ friction	0.60±0.04	0.57±0.03	0.12±0.02	0.05±0.00	0.54±0.05	0.49±0.02	0.38±0.02	0.40±0.01	0.25±0.07	0.04±0.01
Charged two-body (Coulomb)	0.41±0.04	0.22±0.06	0.19±0.03	0.03±0.00	0.40±0.06	0.23±0.04	0.46±0.02	0.42±0.02	0.49±0.11	0.02±0.01

H.2 THREE-BODY OOD BY DYNAMICAL REGIME

We bin the held-out three-body test trajectories by minimum inter-body distance over the clip (bin 0 = closest encounter / most chaotic; bin 2 = widest / most regular) and recompute the map-free metrics per regime (Table 48). The auxiliary model's local (mutual-kNN) advantage is present in every regime but scales with regularity: it is largest in the regular regime (0.17 vs. 0.07) and shrinks, without vanishing, under chaotic close encounters (0.11 vs. 0.09) so the induced correspondence transfers best to near-regular OOD dynamics and degrades under chaos. We omit per-bin CKA here: on these small, homogeneous strata map-free CKA is dominated by within-bin covariance shrinkage rather than by cross-modal correspondence, so the regime analysis rests on the identity-level mutual-kNN, the metric that isolates the per-trajectory correspondence.

Table 48: Three-body OOD stratified by regime (min inter-body distance; bin 0 = most chaotic). The auxiliary local-neighborhood advantage is chaos-dependent.

regime	aux mkNN	no-aux mkNN
chaotic (bin 0)	0.11±0.02	0.09±0.01
mid (bin 1)	0.15±0.02	0.09±0.01
regular (bin 2)	0.17±0.03	0.07±0.00

I WHY ALIGNMENT REQUIRES A REFERENCE: AN ANALYSIS

This appendix states and proves the results underlying §4 (§I.1) and gives the supporting analyses (§I.2–§I.4). Throughout, $e = f(x) \in E$ and $e' = g(y) \in E'$ are the two latents under the coupling induced by h, with finite second moments, centered, with nonsingular covariances $\Sigma_V = \mathrm{Cov}(e)$, $\Sigma_T = \mathrm{Cov}(e')$ and cross-covariance $\Sigma_{VT} = \mathrm{Cov}(e, e')$, and $\mu_f(h) := \mathbb{E}[f(x) \mid h]$, $\mu_g(h) := \mathbb{E}[g(y) \mid h]$ are the conditional means.

I.1 PROOFS FOR §4

Probe–subspace equivalence. For any encoder f and any affine π, orthogonality of conditional expectation gives $\mathbb{E}\|\pi(f(x)) - s(h)\|_2^2 = \mathbb{E}\|\pi(f(x)) - s^*(x)\|_2^2 + \mathbb{E}\|s^*(x) - s(h)\|_2^2$, since $\pi(f(x)) - s^*(x)$ is a function of x and $\mathbb{E}[s(h) \mid x] = s^*(x)$. Hence $\mathrm{dec}_H(f)$ never falls below the second term, the irreducible error, and attains it exactly when some affine π recovers $s^*(x)$ almost surely. Assume $\mathrm{Cov}(s^*)$ is nonsingular and $d_h \le d$. If $\pi(e) = Ae + b$ attains the irreducible error, then $\pi(f(x)) = s^*(x)$ almost surely; since $\mathrm{Cov}(s^*) = A\,\Sigma_V\,A^\top$ is nonsingular, $A \in \mathbb{R}^{d_h \times d}$ has full row rank, so $\dim\ker A = d - d_h$. Take $\mathcal{N} = \ker A$ and any complement $\mathcal{S}$ of dimension d_h: $A|_{\mathcal{S}}$ is invertible, and the $\mathcal{S}$-component of $f(x)$ equals $(A|_{\mathcal{S}})^{-1}\big(s^*(x) - b\big)$, an affine bijection of $s^*(x)$, so the information subspace exists. Conversely, if the subspace exists, composing the inverse of the

bijection with the projection along $\mathcal{N}$ gives an affine map sending $f(x)$ to $s^*(x)$, which attains the irreducible error. The solutions of (P^H) are therefore exactly the H-sufficient encoders.

Definition 3 (Cross-modal transport and cycle operator)**.** The *transport maps* are the population cross-modal regressions

$$W^*_{V\to T} \;=\; \arg\min_{W}\; \mathbb{E}\,\|e' - We\|_2^2 \;=\; \Sigma_{TV}\Sigma_V^{-1}, \qquad W^*_{T\to V} \;=\; \Sigma_{VT}\Sigma_T^{-1}, \tag{21}$$

and the *cycle operator* of the video side is

$$C_V \;=\; W^*_{T\to V}\, W^*_{V\to T}. \tag{22}$$

Each map is determined by its own regression problem, so the cycle measures how much of the latent is preserved under transport to the other space and back; §6.2 fits the maps the same way.

Lemma 4 (The cycle spectrum is the squared canonical correlations)**.** *Let $K = \Sigma_V^{-1/2}\Sigma_{VT}\Sigma_T^{-1/2}$, whose singular values $1 \ge \rho_1 \ge \cdots \ge \rho_d \ge 0$ are the canonical correlations of (e, e') (Hotelling, 1936). Then: (i) $\Sigma_V^{-1/2} C_V\, \Sigma_V^{1/2} = KK^\top$, so C_V is similar to a symmetric positive-semidefinite matrix with eigenvalues $\rho_1^2, \ldots, \rho_d^2$; (ii) under the fork of Eq.* (1)*, $\Sigma_{VT} = \mathrm{Cov}\big(\mu_f(h),\, \mu_g(h)\big)$: all linear cross-modal transport flows through the conditional means.*

Proof. (i) Substituting Eq. (21), $C_V \;=\; \Sigma_{VT}\Sigma_T^{-1}\Sigma_{TV}\Sigma_V^{-1}$; conjugating by $\Sigma_V^{-1/2}$ gives $\Sigma_V^{-1/2}\Sigma_{VT}\Sigma_T^{-1/2}\cdot\Sigma_T^{-1/2}\Sigma_{TV}\Sigma_V^{-1/2} = KK^\top$. (ii) The law of total covariance decomposes the cross-covariance as

$$\Sigma_{VT} \;=\; \mathbb{E}\big[\mathrm{Cov}(e, e' \mid h)\big] \;+\; \mathrm{Cov}\big(\mathbb{E}[e \mid h],\, \mathbb{E}[e' \mid h]\big),$$

and the first term vanishes because $e \perp e' \mid h$ (the fork of Eq. (1), inherited by the deterministic encoders; §3), leaving $\Sigma_{VT} = \mathrm{Cov}\big(\mu_f(h), \mu_g(h)\big)$. □

Proposition 5 (Unit correlation characterizes a shared statistic of h)**.** *In the setting of Definition 3:*

- *(i)* (Characterization.) *For $k \ge 1$, $\rho_1 = \cdots = \rho_k = 1$ if and only if there exist linear maps $U, V : \mathbb{R}^d \to \mathbb{R}^k$ of rank k and a statistic $c : \mathcal{H} \to \mathbb{R}^k$ with $Ue = Ve' = c(h)$ almost surely. Directions of unit correlation are exactly the linear statistics of h carried noiselessly by both latents, i.e. a shared reference.*

- *(ii)* (Restricted regime.) *If in addition the pair has the affine form $e = R_V s(h) + \varepsilon_V$, $e' = R_T s(h) + \varepsilon_T$ (Eq.* (24)*) with $\Sigma_s \succ 0$, R_V, R_T of full column rank d_h, and residuals satisfying $\mathbb{E}[\varepsilon \mid h] = 0$, supported on the orthogonal complements of the signal subspaces with nonsingular covariance there (so $\Sigma_V, \Sigma_T \succ 0$), then*

$$C_V \;=\; \Pi_{\mathcal{S}_V},$$

the orthogonal *projection onto $\mathcal{S}_V = \mathrm{col}(R_V)$: the round trip returns every reference-determined coordinate exactly, acting as the identity on the H-subspace, and annihilates the complement. Under these assumptions, $s(h)$ is linearly recoverable from the signal component of the latent; under partial observability the unit block becomes a diagonal of attenuation factors $\rho_i^2 < 1$ (Spearman, 1904), a shrunken projection, and the identity clause holds on the components both observations determine. In particular $C_V \neq I$ whenever $d_h < d$ (indeed $\|C_V - I\|_F / \|I\|_F = \sqrt{(d-d_h)/d}$), so perfect cycle retrieval coexists with a cycle matrix far from the identity (Table 17; App. I.3).*

Proof of Proposition 5. (i) combines the classical equivalence of unit correlation with a common random variable (Rényi, 1959; Gács & Körner, 1973; Witsenhausen, 1975) with the fork (Gresele et al., 2019; von Kügelgen et al., 2021). ($\Leftarrow$) Since U has rank k and $\Sigma_V \succ 0$, $\mathrm{Cov}(c) = U\Sigma_V U^\top \succ 0$; replacing (U, V, c) by $(\mathrm{Cov}(c)^{-1/2}U,\ \mathrm{Cov}(c)^{-1/2}V,\ \mathrm{Cov}(c)^{-1/2}c)$ we may take $\mathrm{Cov}(c) = I_k$. Let $\alpha_i = \Sigma_V^{1/2}U_i^\top$ and $\beta_i = \Sigma_T^{1/2}V_i^\top$, where U_i, V_i are the rows of U, V. Since $U\Sigma_V U^\top = V\Sigma_T V^\top = \mathrm{Cov}(c) = I_k$, each family is orthonormal, and $\alpha_i^\top K\beta_j = U_i\Sigma_{VT}V_j^\top = \mathrm{Cov}(c_i, c_j) = \delta_{ij}$. As $\|K\|_2 = \rho_1 \le 1$ and $\alpha_i^\top K\beta_i = 1$ for unit vectors, equality in Cauchy–Schwarz gives $K\beta_i = \alpha_i$ and $K^\top\alpha_i = \beta_i$, so $K^\top K\beta_i = \beta_i$: the k orthonormal β_i are right singular vectors of

K with singular value 1, hence $\rho_1 = \cdots = \rho_k = 1$. ($\Rightarrow$) Let $(u_i, v_i) = (a_i^\top e, b_i^\top e')$, $i \le k$, be canonical pairs, centered with unit variance and $\mathrm{corr}(u_i, v_i) = 1$, so $\mathbb{E}[(u_i - v_i)^2] = 2 - 2\rho_i = 0$ and $u_i = v_i$ a.s. Since u_i is a function of e and v_i of e', they are conditionally independent given h; a variable a.s. equal to one conditionally independent of it satisfies $\mathrm{Var}(u_i \mid h) = \mathrm{Cov}(u_i, v_i \mid h) = 0$ a.s., so $u_i = \mathbb{E}[u_i \mid h] =: c_i(h)$ a.s. The canonical directions $a_1, \ldots, a_k$ are linearly independent (their images are uncorrelated with unit variance and $\Sigma_V \succ 0$), likewise the b_i; stacking gives U, V of rank k and $c = (c_1, \ldots, c_k)$. (ii) By $\mathbb{E}[\varepsilon \mid h] = 0$ the conditional means are $\mu_f(h) = R_V s(h)$ and $\mu_g(h) = R_T s(h)$, so Lemma 4(ii) gives $\Sigma_{VT} = R_V \Sigma_s R_T^\top$, supported on $\mathcal{S}_V \times \mathcal{S}_T$; the residual support condition and $\mathbb{E}[s(h)\,\varepsilon^\top] = 0$ make Σ_V, Σ_T block-diagonal in orthonormal bases adapted to $\mathcal{S}_V \oplus \mathcal{S}_V^\perp$ and $\mathcal{S}_T \oplus \mathcal{S}_T^\perp$. Writing $\tilde{R}_V, \tilde{R}_T \in \mathbb{R}^{d_h \times d_h}$ (invertible, by full column rank) for the frames in these bases, the signal block of C_V is $(\tilde{R}_V \Sigma_s \tilde{R}_T^\top)(\tilde{R}_T \Sigma_s \tilde{R}_T^\top)^{-1}(\tilde{R}_T \Sigma_s \tilde{R}_V^\top)(\tilde{R}_V \Sigma_s \tilde{R}_V^\top)^{-1} = I_{d_h}$, and every other block vanishes; hence $C_V = \Pi_{\mathcal{S}_V}$, whose distance from the identity is $\|I - \Pi_{\mathcal{S}_V}\|_F = \sqrt{d - d_h}$. (With in-signal residual noise the unit block becomes a diagonal of attenuation factors, a shrunken projection; this is Spearman's attenuation (Spearman, 1904).) □

Corollary 6 (Subspace unitarity of the transport)**.** *Under the hypotheses of Proposition 5(ii), in the adapted bases of its proof: (i) $W^*_{V\to T} = \tilde{R}_T \tilde{R}_V^{-1} \oplus 0$: the transport acts on the signal subspace as the frame change $R_T R_V^+$ of Eq.* (25) *and annihilates the complement; it is a general linear map, generically non-orthogonal. (ii) In whitened coordinates the transport is a partial isometry: the signal block K_s of $K = \Sigma_V^{-1/2} \Sigma_{VT} \Sigma_T^{-1/2}$ is an orthogonal matrix, and K vanishes off it.*

Proof of Corollary 6. (i) With Σ_{VT} and Σ_V block-structured as in the proof of Proposition 5(ii), $W^*_{V\to T} = \Sigma_{TV}\Sigma_V^{-1}$ has signal block $\tilde{R}_T \Sigma_s \tilde{R}_V^\top (\tilde{R}_V \Sigma_s \tilde{R}_V^\top)^{-1} = \tilde{R}_T \tilde{R}_V^{-1}$ and vanishes elsewhere. (ii) By Lemma 4(i), $KK^\top$ is similar to C_V by a block-diagonal conjugation, so its signal block equals I_{d_h}, and K vanishes off the signal blocks; a square block with $K_s K_s^\top = I$ is orthogonal, so the canonical variates of the two blocks are related by an orthogonal transformation, the standard CCA geometry (Anderson, 2003). □

Theorem 7 (the H-subspace)**.** *In the setting of Definition 3, let $1 \ge \rho_1 \ge \cdots \ge \rho_d \ge 0$ be the canonical correlations of e and e'. Then:*

- *(i)* (Spectrum.) *C_V is similar to $\mathrm{diag}(\rho_1^2, \ldots, \rho_d^2)$: the round trip acts as the identity on its eigenspace of eigenvalue 1, and every eigendirection with $\rho_i < 1$ is scaled by ρ_i^2.*
- *(ii)* (Characterization.) *That eigenspace is the H-subspace $\mathcal{S}_V$ of Definition 1 (with c taken maximal), and its dimension k is the number of unit canonical correlations: the subspace on which the two modalities agree is spanned by the loading directions of the linear statistics of h present in both latents.*
- *(iii)* (Geometry.) *Restricted to the H-subspace the transport $W^*_{V\to T}$ is a linear bijection onto $\mathcal{S}_T$ and, in whitened coordinates, a linear isometry; in raw coordinates it is linear but in general not orthogonal, since orthogonality holds only in the whitened frame.*

Proof. (i) is Lemma 4(i); a matrix similar to a symmetric one is diagonalizable with the same eigenvalues. (ii) Let m be the number of unit canonical correlations. By Proposition 5(i, $\Rightarrow$) the corresponding canonical pairs give a shared statistic of dimension m, so the maximal c of Definition 1 has $k \ge m$; by Proposition 5(i, $\Leftarrow$) a shared statistic of dimension k forces k unit correlations, so $k \le m$. Hence $k = m$, and we may take $c_i(h) = a_i^\top e = b_i^\top e'$ with (a_i, b_i) the unit-correlation canonical directions. The vectors $\Sigma_V a_i = \mathrm{Cov}(e, c_i)$ are eigenvectors of eigenvalue 1: using $a_i^\top e = b_i^\top e'$ almost surely, $C_V \Sigma_V a_i = \Sigma_{VT}\Sigma_T^{-1}\Sigma_{TV} a_i = \Sigma_{VT}\Sigma_T^{-1}\,\mathrm{Cov}(e', b_i^\top e') = \Sigma_{VT} b_i = \mathrm{Cov}(e, a_i^\top e) = \Sigma_V a_i$. They are linearly independent because the a_i are and $\Sigma_V \succ 0$, and by (i) the eigenvalue 1 has multiplicity exactly $m = k$, so they span the whole eigenspace, which is therefore $\mathrm{col}\,\mathrm{Cov}(e, c) = \mathcal{S}_V$. (iii) The same computation gives $W^*_{V\to T}\Sigma_V a_i = \Sigma_{TV} a_i = \Sigma_T b_i$, and the $\Sigma_T b_i = \mathrm{Cov}(e', c_i)$ form a basis of $\mathcal{S}_T$, so the transport carries a basis of $\mathcal{S}_V$ onto a basis of $\mathcal{S}_T$. In whitened coordinates the transport is $\Sigma_T^{-1/2} W^*_{V\to T} \Sigma_V^{1/2} = K^\top$, whose singular values on the unit-correlation block are all 1, hence a linear isometry there; Corollary 6 gives the raw-coordinate form $\tilde{R}_T \tilde{R}_V^{-1}$ under Eq. (24). □

Nothing in Theorem 7 is assumed of the models: the fork holds by construction of the dataset (Eq. (1)), finite second moments and nonsingularity can be checked on the trained latents, and the statements apply to any pair of encoders, trained or not.

I.2 No shared frame without a reference

An encoder trained solely to model the distribution of its own modality is largely unconstrained in its choice of coordinate frame. For a decoder that can absorb an invertible re-parameterization of its input (a linear first layer absorbs any element of $\mathrm{GL}(d)$), the training objective is unchanged by that re-parameterization. Two encoders trained separately on different modalities therefore each occupy one frame from among a continuum of equally optimal ones, and independently chosen frames should not coincide, even when both latents capture the underlying physics.

To make this precise within this idealization, let f and g be optimal encoders for the two modalities. Because each objective is invariant under invertible re-parameterizations the decoder can absorb, its set of optimal solutions is closed under $\mathrm{GL}(d)$: the frame, and with it the map $h \mapsto \mathbb{E}[e \mid h]$ from cause to conditional mean, is free. Which statistics of h a latent carries is fixed by the data, not the frame, and a statistic both latents carry is shared in any pair of frames: $\Sigma_{VT} = \mathrm{Cov}(\mu_f(h), \mu_g(h))$ (Lemma 4(ii)) is nonzero exactly when some linear read-out of $\mu_f(h)$ is correlated with some linear read-out of $\mu_g(h)$, and no common frame is required for such shared content to exist. This is the content Definition 1 counts and canonical correlation measures, and the $\lambda{=}0$ latents have some of it: the two near-unit directions of Table 32 are approximately statistics of h that both modalities display without any target (permutation null ≈ 0). They place a trajectory among the right neighbours without fixing a common frame, which is why neighbourhood agreement sits above its null (mutual-kNN 7$\times$ chance) while per-trajectory retrieval stays at the floor.

Alignment in the sense of Eq. (5) asks for more than shared content. It requires a single modality-agnostic $\mu : \mathcal{H} \to \mathbb{R}^d$ and one transform T with $\mathbb{E}[e \mid h] \approx \mu(h)$ and $\mathbb{E}[e' \mid h] \approx T^{-1}\mu(h)$ for all h at once: the two conditional-mean geometries must be the same function of the cause up to one fixed T, on every direction. A marginal objective ties $h \mapsto \mathbb{E}[e \mid h]$ to nothing in the other modality; the two maps are chosen independently within their sets of optima, so on every direction the shared content does not fix, nothing ties them to a common (μ, T), and Eq. (5) holds only on that content (Table 18). This frame freedom accounts for the map-free metrics; a fitted map absorbs any frame, and fails at $\lambda{=}0$ because two shared directions are too few for reliable identification (1.16 bits by the KSG estimate; Fig. 8, App. D.4).

The agreement of independently chosen frames is also sharply concentrated in high dimension. Consider the null in which the two frames differ, on a fixed direction, by a uniformly random rotation $Q \sim \mathrm{Unif}(\mathrm{SO}(d))$, a frame chosen without reference to the other modality. For any fixed unit direction u and $d \geq 2$, the vector Qu is uniformly distributed on the sphere S^{d-1}, and the standard spherical-cap estimate gives $\Pr[\,|\langle u, Qu\rangle| \geq t\,] \leq 2e^{-c\,d\,t^2}$ (Vershynin, 2018) for a universal constant $c > 0$. For the squared-cosine statistic $S_d = \langle u, Qu\rangle^2$, whose mean is $\mathbb{E}S_d = 1/d$, this yields, for any $\epsilon > 0$,

$$\Pr\big[S_d \geq \epsilon\big] \;\leq\; 2\,e^{-c\,d\,\epsilon}, \tag{23}$$

an exponentially small probability in the latent dimension. Frame agreement between independently chosen frames is therefore vanishing in expectation and concentrated: in high dimension, two such frames agree to within ϵ on a fixed direction only with exponentially small probability. Alignment beyond what shared content supplies must be attributed to a mechanism that selects a common frame. §I.4 identifies it: the reference constrains both encoders to a common $s(h)$, at which point a single $T = R_T R_V^{+}$ works by construction on the target-relevant subspace.

I.3 The correspondence lives in a low-dimensional shared subspace

The reference does more than align the two spaces: the correspondence it induces between them is carried within a low-dimensional subspace of the 384-dimensional ambient space, even though neither space is itself low-dimensional.

First, the round-trip geometry of §6.2 separates a faithful low-dimensional subspace from an unconstrained complement. The cycle $W_{v\to t}W_{t\to v}$ returns each point to itself for retrieval (cycle recall-at-1 $\approx$ 0.98–1.00), yet its matrix is far from the identity globally ($\|WW - I\|/\|I\| \approx 1.65$). Both facts

follow if the composition is the identity *on* the subspace that carries the cross-modal signal (the H-subspace of Theorem 7) and unconstrained *off* it. Under the idealization of Proposition 5(ii), the population cycle annihilates that complement, which would give $\|C_V - I\|_F / \|I\|_F = \sqrt{(d-k)/d} < 1$; the fitted finite-sample maps are not constrained by any signal there, which is why the measured deviation exceeds this bound.

Second, this low dimension is *shared* across the two modalities. Under the reference the two latents need a similar number of dimensions to explain 90% of their variance (12.9 vision, 12.3 text), whereas without it their participation ratios differ by $\sim 2.8\times$ (Table 24). The reference ties both modalities to the same low-dimensional structure instead of relating two full-dimensional spaces, and the cross-modal map inherits that low dimension, acting only where both spaces carry the shared signal.

I.4 Alignment with a reference: removing the relative frame freedom

§I.2 concerns *marginal* objectives: with no term referencing h, the frame is free and frame agreement is concentrated near zero. The auxiliary objective gives both encoders a reason to encode the target; under the idealization below, this removes the relative frame freedom on the target-relevant subspace.

The reference fixes the geometry on the target-relevant subspace. Adding $\mathcal{L}_{\text{aux}}$ (Eq. (3)) makes the target linearly decodable from each latent at the optimum of the reference problem (§I.1); it does not by itself fix the complete latent frame. To characterize the geometry this induces when both encoders expose the same reference, we analyze the following idealization, motivated by linear identifiability (Hyvärinen & Morioka, 2016; Khemakhem et al., 2020; Roeder et al., 2021; Reizinger et al., 2025): suppose the target-relevant component of each latent factors through a common statistic $s(h) \in \mathbb{R}^{d_h}$,

$$f(x) \;=\; R_V\, s(h) \;+\; \varepsilon_V, \qquad g(y) \;=\; R_T\, s(h) \;+\; \varepsilon_T, \tag{24}$$

with modality-specific frames $R_V, R_T \in \mathbb{R}^{d \times d_h}$ of full column rank and residuals $\varepsilon_V, \varepsilon_T$ orthogonal to the signal subspaces $\mathcal{S}_V = \mathrm{col}(R_V)$, $\mathcal{S}_T = \mathrm{col}(R_T)$. Under this structure, with the conditions of Proposition 5(ii), $\mathcal{S}_V, \mathcal{S}_T$ are the H-subspace of Definition 1, and the two frames differ only in how each embeds the shared statistic $s(h)$. Everything below is conditional on Eq. (24); the experiments test its predicted consequences, not its premises.

The relative frame freedom is removed. Where the marginal objective left $T_{V \to T}$ free, Eq. (24) pins it on the source signal subspace $\mathcal{S}_V$:

$$T_{V \to T}\big|_{\mathcal{S}_V} \;=\; R_T R_V^{+}, \tag{25}$$

since $R_T R_V^{+} R_V s(h) = R_T s(h)$ with $R_V^{+} R_V = I_{d_h}$. The map is determined by the two frames' relation to the common $s(h)$, so a single linear map relates the two codes on $\mathcal{S}_V$ by construction, without relying on two independently chosen frames to agree. In the text→vision direction this is the map $W \approx R_V R_T^{+}$ observed in §6.2: a general linear map, orthogonal on the signal subspace only when the two frames differ by a rotation (as when both are whitened), which our unnormalized objectives do not impose. On the complement of $\mathcal{S}_V$ the map remains free, as in §I.2, so the reference both induces alignment and confines it to the image of $s(h)$; App. H tests how much of this correspondence transfers when the dynamics change.

The asymmetric case: alignment bounded by shared information *and* target rank. The asymmetric-target ablation (Table 28) breaks the common-reference assumption: vision regresses the scalar energy $s_V(h) = \mathcal{E}(\theta, q_0, p_0)$ and text regresses the parameter vector $s_T(h) = \theta$, with no shared component. Under the analogous factorization, $f(x) = R_V\, s_V(h) + \varepsilon_V$ and $g(y) = R_T\, s_T(h) + \varepsilon_T$, so the best linear map from the text code to the vision code is the population regression of $f(x)$ on $g(y)$, which relates embeddings of two different functions of h, instead of the shared-frame map $R_V R_T^{+}$. Two effects cap it. First, *content*: its R^2 is governed by how much s_V and s_T co-vary under the latent prior; for a scalar s_V and noiseless signals, its R^2 along the vision signal direction is the squared single canonical correlation ρ_1^2 between $\mathcal{E}$ and θ (the squared multiple correlation of energy on the parameters), with residual noise attenuating it further. It is positive whenever $\mathcal{E}$ is linearly correlated with θ, and $\rho_1 = 1$ if and only if $\mathcal{E}$ is almost surely an affine function of θ. Second, *rank*: because s_V is scalar, the vision signal is one-dimensional, so only a one-dimensional

component of θ is recoverable, a thinner correspondence independent of ρ_1; consistent with this, the asymmetric target transfers less out of distribution (Table 46). Both effects predict the partial alignment we observe: the asymmetric pair aligns well above the $\lambda{=}0$ floor but below the identical target (CKA 0.41 vs. 0.72; Table 28). The identical target is the limiting case $s_V = s_T = s$, in which the full frame is shared on $\mathcal{S}_V$ and the map of Eq. (25) is recovered. In Definition 1's terms, what can be shared is a common statistic $c(h)$, not necessarily an identical target $s(h)$.