

Platonic Projection Structures: Operator-Induced Observability in Representation Learning

Kazuo Ishii

kazuoishii2014@gmail.com

Suwa University of Science <https://orcid.org/0000-0002-8363-8266>

Bishnu Prasad Gautam

Suwa University of Science

Jieling Wu

Suwa University of Science

Saher Javaid

Kanazawa Gakuin University

Research Article

Keywords: observability, representation learning, operator theory, positive semidefinite operators, knowledge distillation, interpretability, quotient geometry, explainable AI

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9749304/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Platonic Projection Structures: Operator-Induced Observability in Representation Learning

Kazuo Ishii^{1*}, Bishnu Prasad Gautam^{1†}, Jieling Wu^{1‡} and Saher Javaid^{2§}

¹Department of Applied Information Engineering
Faculty of Engineering
Suwa University of Science, Chino, Nagano, Japan

²Department of Information Engineering
Kanazawa Gakuin University, Kanazawa, Japan

Abstract

We introduce Platonic Projection Structures (PPS), an operator-theoretic framework for analyzing observability in representation learning systems. Rather than treating observable outputs as direct reflections of latent representations, PPS models observation as a geometry induced by a self-adjoint positive semidefinite operator acting on a latent Hilbert space. Within PPS, a system is represented as a triple $(\mathcal{H}, \Pi, O)$, where $\mathcal{H}$ denotes a latent representation space, $\Pi \succeq 0$ is an observation operator, and $O(v) = \langle v, \Pi v \rangle$ defines an induced observable quantity. The framework characterizes observability through the quotient geometry $\mathcal{H}/\ker(\Pi)$, which represents equivalence classes of latent states that are indistinguishable under observation. From this perspective, observable behavior is governed not by latent representations themselves, but by the geometry induced through the observation operator. We show that both quantum measurement and deep learning inference can be formulated within this common operator-theoretic structure while differing in the algebraic properties of their observation operators. Within this perspective, quantum measurement serves primarily as a mathematically canonical example of projection-mediated observability. The correspondence developed in PPS is therefore structural rather than physical. Within the same framework, representation transfer and knowledge distillation can be interpreted as approximate preservation of observable geometry through the intertwining condition $\Phi\Pi_T \approx \Pi_S\Phi$. PPS further reveals a structural limitation of output-based interpretability: latent components contained in $\ker(\Pi)$ are fundamentally inaccessible from observables generated through the induced observation process. Accordingly, attribution and explanation methods inherit intrinsic constraints imposed by the observation geometry itself. We provide controlled empirical validations demonstrating kernel-invariant observability, projection-induced attribution gaps, and rank-controlled observable geometry in latent representation spaces. Overall, PPS provides a mathematically explicit operator-theoretic perspective for analyzing observability, interpretability, and representation transfer in projection-mediated learning systems.

*Corresponding author: kishii@rs.sus.ac.jp

†gautam.bishnu@rs.sus.ac.jp

‡wu.jieling@rs.sus.ac.jp

§saher@kanazawa-gu.ac.jp

1 Introduction

Modern machine learning systems are increasingly interpreted through observable quantities such as output distributions, attribution maps, intermediate representations, and feature activations [1, 2, 3, 4, 5, 6, 7]. At the same time, recent developments in representation learning, interpretability, and information-theoretic analysis have highlighted that observable outputs often provide only partial access to the underlying latent structure of a system. Despite substantial progress in explainable artificial intelligence (XAI), representation analysis, and information bottleneck approaches [8, 9, 10, 11, 1, 2, 3], the notion of *observability* itself remains only partially formalized in modern learning systems. In many existing formulations, observable outputs are implicitly treated as sufficiently informative reflections of latent representations. However, practical learning systems frequently operate under restricted observation geometries in which only selected latent directions contribute to observable behavior. This paper introduces *Platonic Projection Structures* (PPS), an operator-theoretic framework for analyzing projection-mediated observability in representation learning systems [12, 13, 14]. Rather than treating observation as direct access to latent states, PPS models observable quantities as induced through a self-adjoint positive semidefinite operator acting on a latent Hilbert space [12, 14]. Within PPS, a system is represented as

$$(\mathcal{H}, \Pi, O), \quad (1)$$

where $\mathcal{H}$ denotes a latent representation space, $\Pi \succeq 0$ is an observation operator, and the observable quantity is defined by

$$O(v) = \langle v, \Pi v \rangle. \quad (2)$$

The central principle of PPS is that observability is governed not by latent representations themselves, but by the geometry induced through the observation operator. These assumptions are standard in operator theory and positive semidefinite analysis [13, 14].

The operator Π induces the equivalence relation

$$v_1 \sim_{\Pi} v_2 \iff v_1 - v_2 \in \ker(\Pi), \quad (3)$$

yielding the quotient structure

$$\mathcal{H} / \ker(\Pi), \quad (4)$$

which represents the effective observable space. Under this formulation, latent components contained in $\ker(\Pi)$ are structurally inaccessible from observables generated through Eq. (2). This viewpoint leads naturally to a spectral interpretation of observability [15, 16, 17, 18, 19]. The eigenspectrum of Π determines which latent directions are strongly observable, weakly observable, or entirely suppressed. Consequently, the effective observable dimension is governed not by the ambient latent dimensionality itself, but by the rank and spectral structure of the observation operator [18, 15, 20]. The PPS framework provides a unified abstraction for projection-mediated observation across multiple domains. This perspective is related to operator-theoretic formulations appearing in quantum mechanics, kernel methods, and information geometry [21, 22, 23, 24, 25, 26, 20]. In quantum systems, idealized measurements correspond to orthogonal projections satisfying [21, 22]

$$\Pi^2 = \Pi, \quad (5)$$

whereas deep learning systems generally induce non-idempotent positive semidefinite operators. Despite these differences, both settings share the same abstract operator-theoretic structure: observable quantities are generated through operator-induced structure rather than through direct

access to latent states. Within this perspective, representation transfer and knowledge distillation can also be interpreted geometrically. Within PPS, projection-mediated observability is treated as an explicit mathematical structure rather than as a purely metaphorical analogy between domains.

Given teacher and student observation operators Π_T and Π_S , a transfer map

$$\Phi : \mathcal{H}_T \rightarrow \mathcal{H}_S$$

approximately preserves observable geometry when

$$\Phi \Pi_T \approx \Pi_S \Phi. \quad (6)$$

This formulation is motivated by approximate intertwining relations commonly appearing in operator-theoretic alignment and representation-transfer settings. Equation (6) interprets distillation as approximate consistency between induced observation geometries rather than solely as output-level distribution matching [27, 28, 29, 30, 31]. A central implication of PPS concerns the structural limits of output-based interpretability. Since observables generated through Eq. (2) are invariant under perturbations contained in $\ker(\Pi)$, they depend only on the equivalence classes induced by Eq. (3) and represented by the quotient structure Eq. (4). Latent components lying in $\ker(\Pi)$ cannot, in general, be recovered from observables alone. From this perspective, limitations of interpretability are not merely algorithmic, but geometric and operator-induced.

The PPS framework is conceptually related to several existing research directions. Information bottleneck formulations characterize information compression through mutual-information constraints [32, 8, 9, 10, 11], whereas PPS characterizes observability through operator-induced geometry on latent spaces. Similarly, concept bottleneck models [33] and representation-constrained learning methods may be interpreted as imposing explicit structural constraints on induced observable subspaces [33, 34, 35, 36]. PPS also exhibits structural parallels with observability theory in control systems [37], although the present framework focuses on projection-induced accessibility in representation spaces rather than dynamical state reconstruction. To empirically examine the PPS formulation, we construct controlled latent-space experiments directly evaluating projection-induced observability. The experiments verify three central predictions of the framework: (i) observables generated through Eq. (2) remain invariant under perturbations confined to $\ker(\Pi)$, (ii) the rank structure of Π governs effective observable dimensionality and predictive behavior, and (iii) attribution methods inherit structural limitations induced by the observation geometry.

The primary contribution of this work is conceptual rather than architectural. PPS provides a unified operator-theoretic perspective on observability, interpretability, and representation transfer in projection-mediated learning systems. By emphasizing operator-induced geometry, the framework shifts attention from observable outputs alone to the structural mechanisms determining what can become observable in the first place [20, 14]. Importantly, PPS does not aim to introduce a new class of operators or geometric spaces. Rather, its novelty lies in treating observability itself as a primitive operator-induced geometric structure governing representation accessibility. In contrast to kernel similarity methods or probabilistic information-geometric formulations, PPS explicitly centers the quotient structure induced by $\ker(\Pi)$ as the defining object of observable equivalence. PPS is therefore intended primarily as a structural framework for reasoning about observability across representation systems.

The resulting framework provides a mathematically explicit perspective on observability across projection-mediated representation systems. The main contributions of this paper are summarized as follows:

1. We introduce PPS, an operator-theoretic framework for projection-mediated observability in representation learning systems.

2. We characterize observability through quotient geometry and the spectral structure of positive semidefinite observation operators.
3. We formulate representation transfer as approximate preservation of observable geometry via the operator-consistency relation in Eq. (6).
4. We analyze structural limitations of output-based interpretability arising from the kernel structure of observation operators.
5. We provide controlled empirical analyses of kernel-invariant observability, projection-induced attribution gaps, and rank-controlled observable geometry.

2 Operator-Theoretic Framework of Observability

This section introduces Platonic Projection Structures (PPS) as an operator-theoretic framework for analyzing projection-mediated observability in representation learning systems. Rather than treating observable quantities as direct reflections of latent states, PPS models observation as a geometric process induced through a self-adjoint positive semidefinite operator acting on a latent Hilbert space.

2.1 Platonic Projection Structures

Let $\mathcal{H}$ denote a real or complex Hilbert space of latent representations. Within the PPS framework introduced in Eq. (1), observability is induced through an observation operator $\Pi : \mathcal{H} \rightarrow \mathcal{H}$, which generates observable quantities through the quadratic functional defined in Eq. (2). We distinguish two regimes of observation:

1. strict projection operators satisfying Eq. (5),
2. generalized positive semidefinite operators satisfying $\Pi \succeq 0$.

Idealized quantum measurements naturally belong to the first regime [21, 22], whereas representation learning systems are generally associated with the second. Despite this distinction, both settings share the same structural mechanism: observable quantities are generated through operator-induced structure rather than through direct access to latent states. The observation operator is assumed to satisfy

$$\Pi^\dagger = \Pi, \quad \langle v, \Pi v \rangle \geq 0 \quad \forall v \in \mathcal{H},$$

ensuring that the observable functional in Eq. (2) is well-defined and non-negative. Under PPS, the operator Π becomes the primitive object governing observability. Observable structure is therefore determined not by latent representations themselves, but by the geometry induced through the observation operator.

2.2 Observable Geometry and Quotient Structure

Throughout this paper, the term “observable geometry” refers to the quotient-semimetric structure induced by Π through the equivalence relation in Eq. (3) and the semi-inner product defined in Eq. (7). A central consequence of PPS is that observability is defined only up to the equivalence relation introduced in Eq. (3). Two latent states differing only by a component contained in $\ker(\Pi)$ are observationally indistinguishable under Eq. (2). Accordingly, the effective observable space is not the ambient latent space $\mathcal{H}$ itself, but the quotient geometry defined in Eq. (4) [12, 13]. This

quotient space represents equivalence classes of latent states that generate identical observables under the induced observation process. The operator Π further induces the bilinear form

$$\langle v, w \rangle_{\Pi} := \langle v, \Pi w \rangle, \quad (7)$$

which defines a semi-inner product structure on $\mathcal{H}$. Since directions contained in $\ker(\Pi)$ contribute no observable signal, the form in Eq. (7) becomes non-degenerate only after quotienting by $\ker(\Pi)$. The quotient geometry in Eq. (4) therefore defines the effective observable geometry induced by the observation operator [12, 13]. This establishes a central principle of PPS:

Observable structure is determined not by latent states themselves, but by the geometry induced through the observation operator.

Under this viewpoint, observability becomes a structural property of induced operator geometry rather than a property of individual latent vectors.

2.3 Spectral Structure of Observability

The PPS framework naturally admits a spectral interpretation of observability [15, 16, 17, 18, 19]. Let

$$\Pi = \sum_i \lambda_i e_i e_i^* \quad (8)$$

denote the spectral decomposition of the observation operator. The eigenvalues λ_i determine the degree of observability associated with each latent direction:

$$\lambda_i \gg 0 \Rightarrow \text{strongly observable direction,}$$

$$\lambda_i \approx 0 \Rightarrow \text{weakly observable direction,}$$

$$\lambda_i = 0 \Rightarrow e_i \in \ker(\Pi) \Rightarrow \text{structurally unobservable direction.}$$

Accordingly, the effective observable dimension is governed by

$$\text{rank}(\Pi). \quad (9)$$

Equation (9) implies that observable dimensionality depends not on the ambient latent dimension itself, but on the spectral structure of the observation operator [38, 18]. From an information-theoretic perspective, the eigenspectrum in Eq. (8) characterizes the relative accessibility of latent information under the induced observation process [32, 8, 20]. Large eigenvalues correspond to latent directions strongly represented in observable quantities, whereas directions associated with vanishing eigenvalues become inaccessible through Eq. (2). This spectral viewpoint plays a central role in later experiments, where controlled manipulation of the eigenspectrum systematically alters observable geometry, attribution behavior, and predictive performance.

2.4 Structural Factorization of Observation

The observation process induced by PPS factors naturally through the quotient geometry defined in Eq. (4) [12, 13]. Specifically, the observable functional in Eq. (2) admits the factorization

$$\mathcal{H} \xrightarrow{q} \mathcal{H}/\ker(\Pi) \xrightarrow{\tilde{O}} \mathbb{R}, \quad (10)$$

where q denotes the canonical quotient map. Equation (10) formalizes a central implication of PPS: observable quantities do not uniquely determine latent states, but only equivalence classes induced by the observation geometry. Consequently, latent components lying in $\ker(\Pi)$ are structurally inaccessible from observables generated through Eq. (2). This limitation is therefore geometric rather than merely algorithmic [2, 3]. From an information-geometric viewpoint, the quotient structure in Eq. (4) may be interpreted as the effective observable information geometry induced by the observation operator [4, 5, 6, 3]. The observation process therefore acts as a structured projection mechanism that compresses latent information according to the spectral accessibility encoded in Eq. (8) [8, 10, 11]. Overall, PPS provides a unified operator-theoretic framework in which observability, representation accessibility, and projection-induced information loss are described through induced observable geometry [14, 20, 1].

3 Projection-Mediated Observation Across Systems

This section shows that both quantum measurement and deep learning inference can be interpreted within the PPS framework as instances of projection-mediated observation [21, 22, 24, 23]. The correspondence is structural rather than physical: in both settings, observable quantities are generated through operator-induced geometry acting on latent representation spaces [13, 14, 20].

3.1 Quantum Measurement as Projection-Induced Observation

Quantum measurement provides a canonical example of projection-mediated observability [21, 22]. A quantum system is represented by a normalized state vector

$$|\psi\rangle \in \mathcal{H}_Q, \quad \langle\psi|\psi\rangle = 1,$$

where $\mathcal{H}_Q$ denotes a Hilbert space of admissible states. Measurement is described by a Hermitian operator admitting the spectral decomposition [21, 13]

$$M = \sum_m m P_m, \tag{11}$$

where each P_m is an orthogonal projection operator satisfying Eq. (5). The probability of observing outcome m is given by the Born rule [21, 22]

$$p(m) = \langle\psi|P_m|\psi\rangle. \tag{12}$$

Equation (12) is structurally equivalent to the PPS observable functional defined in Eq. (2) under the identification

$$v = |\psi\rangle, \quad \Pi = P_m.$$

From the PPS viewpoint, quantum measurement can therefore be interpreted as a projection-mediated observation process:

$$|\psi\rangle \rightarrow P_m \rightarrow p(m).$$

This formulation emphasizes that measurement accesses only the projected component of the quantum state. Information contained in complementary directions remains inaccessible under the induced observation process. Importantly, this correspondence is structural rather than physical. PPS does not identify learning systems with quantum systems, but instead uses quantum measurement as a mathematically well-defined example of projection-induced observability [21, 22].

3.2 Deep Learning Inference as Operator-Induced Observation

Deep learning inference can similarly be interpreted within the PPS framework as an operator-mediated observation process [1]. Let

$$f_\theta = f_L \circ f_{L-1} \circ \cdots \circ f_1 \quad (13)$$

denote a deep neural network, and let

$$z = f_\theta(x) \in \mathcal{H}_L$$

represent the final latent representation. In standard architectures, outputs are generated through a linear readout

$$y = Wz, \quad (14)$$

where W maps latent representations to output space. The corresponding quadratic observable becomes [14, 20]

$$\|y\|^2 = \langle z, W^\top W z \rangle. \quad (15)$$

Equation (15) naturally defines the positive semidefinite observation operator [14]

$$\Pi := W^\top W. \quad (16)$$

The operator in Eq. (16) induces the latent observation geometry governing which directions in representation space contribute to observable outputs [15, 18, 20]. Under PPS, deep learning inference can therefore be interpreted as

$$z \rightarrow \Pi \rightarrow \|y\|^2,$$

which is structurally analogous to the projection-mediated process described by Eq. (12). Importantly, PPS does not claim that neural networks are fundamentally quadratic systems. Rather, Eq. (16) identifies a canonical observable geometry induced by the output map. Directions contained in $\ker(\Pi)$ remain inaccessible from observables generated through Eq. (15). Consequently, latent representations are observable only up to the quotient geometry defined in Eq. (4). More generally, Eq. (16) shows that observable behavior in deep learning systems is governed not solely by latent representations themselves, but by the operator-induced geometry through which latent information becomes accessible.

3.3 Structural Correspondence Between Quantum and Learning Systems

Within the PPS framework, both quantum systems and deep learning models can be described through a common abstract structure [21, 24, 23, 25] consisting of:

1. a latent Hilbert space $\mathcal{H}$,
2. an observation operator Π ,
3. an induced observable generated through Eq. (2).

Under this perspective, the two domains admit the formally parallel structure

$$\begin{aligned} \text{Quantum system: } & \mathcal{H}_Q, |\psi\rangle, P_m, p(m) = \langle \psi | P_m | \psi \rangle, \\ \text{Deep learning system: } & \mathcal{H}_L, z, W^\top W, \|y\|^2 = \langle z, W^\top W z \rangle. \end{aligned}$$

In both cases, observable quantities are induced through self-adjoint positive semidefinite operators acting on latent spaces [14, 13]. The distinction lies not in the existence of projection-mediated observability itself, but in the algebraic structure of the corresponding operators. Quantum systems typically involve orthogonal projections satisfying Eq. (5), whereas deep learning systems generally induce non-idempotent positive semidefinite operators such as Eq. (16). This correspondence should therefore be understood as operator-theoretic rather than physical. PPS does not identify neural inference with quantum dynamics, but instead provides a unified geometric language for describing observability across projection-mediated systems [24, 23, 26]. From this viewpoint, observability becomes a property of induced operator geometry rather than a property specific to a particular physical or computational substrate. Figure 1 summarizes this shared projection-mediated structure.

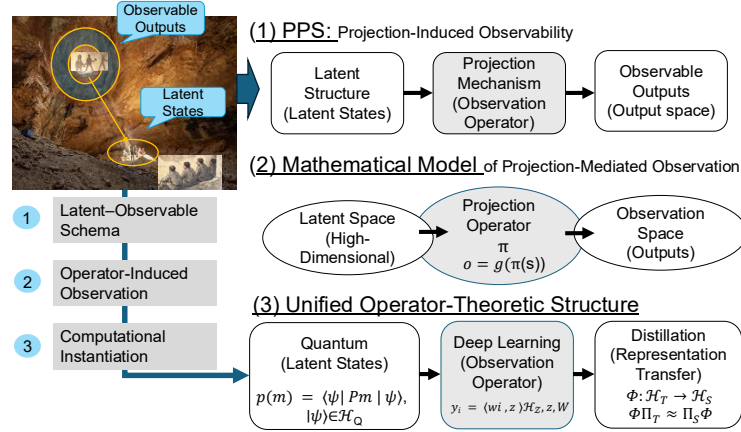


Figure 1: **PPS framework as projection-induced observation geometry.** Schematic illustration of projection-mediated observability under PPS. A latent state $v \in \mathcal{H}$ is mapped through a self-adjoint positive semidefinite operator Π to an observable quantity generated through Eq. (2). Quantum systems correspond to projection operators satisfying Eq. (5), whereas deep learning systems generally induce non-idempotent positive semidefinite operators such as Eq. (16).

4 Information Geometry and Representation Transfer

This section develops an information-theoretic interpretation of PPS and introduces a geometric formulation of representation transfer [32, 8, 20, 27]. Within the PPS framework, observable information is determined not only by latent representations themselves, but by the operator-induced geometry through which latent directions become accessible. From this viewpoint, representation transfer can be interpreted as approximate preservation of observable geometry between latent spaces.

4.1 Information-Theoretic Interpretation of PPS

The PPS framework admits a natural interpretation from the perspective of information accessibility [32, 8, 20]. Let $v \in \mathcal{H}$ be a latent representation, and consider observables generated through the quadratic functional defined in Eq. (2). Under PPS, the observation operator determines which

latent directions contribute to observable quantities. In particular, directions contained in $\ker(\Pi)$ are structurally inaccessible under the induced observation process. The spectral decomposition introduced in Eq. (8) therefore characterizes the relative accessibility of latent information [20, 8]. Large eigenvalues correspond to strongly observable latent directions, whereas vanishing eigenvalues identify directions that are completely unobservable. Accordingly, the effective observable dimension is governed by Eq. (9), rather than by the ambient latent dimensionality itself. From this perspective, the observation process may be interpreted as a structured compression mechanism induced by the observation geometry. Observable quantities generated through Eq. (2) preserve only information contained within the effective observable quotient defined in Eq. (4). This viewpoint is related in spirit to the information bottleneck principle [32, 8, 10, 11], which studies compressed representations through mutual-information constraints. However, PPS differs fundamentally in formulation. Rather than characterizing information through probability distributions over random variables, PPS characterizes observability through the spectral geometry of an operator acting on latent representation spaces [14, 20]. Under PPS, information accessibility is therefore governed geometrically through induced observable structure. The quotient geometry in Eq. (4) may consequently be interpreted as an effective observable information geometry induced by the observation operator.

4.2 Knowledge Distillation as Operator Alignment

Within the PPS framework, knowledge distillation can be interpreted as a geometric alignment problem between operator-induced observable structures [27, 29, 30, 31]. Standard knowledge distillation is commonly formulated by minimizing the divergence between teacher and student output distributions [27]:

$$L_{\text{KD}} = D_{\text{KL}}(p_T \parallel p_S), \quad (17)$$

where

$$p_T = \sigma(W_T z_T), \quad p_S = \sigma(W_S z_S).$$

From the PPS viewpoint, Eq. (17) aligns observable distributions in output space while leaving the geometry of the underlying observation operators implicit. Several extensions, including Fit-Nets [28], RKD [29], and CRD [30], and attention transfer [31], introduce increasingly structured alignment constraints between latent representations. Within PPS, these approaches may be interpreted as partial instances of a broader operator-alignment principle. We associate the following positive semidefinite operators with teacher and student models:

$$\Pi_T = W_T^\top W_T, \quad \Pi_S = W_S^\top W_S. \quad (18)$$

The corresponding quadratic observables become

$$O_T(z_T) = \langle z_T, \Pi_T z_T \rangle, \quad O_S(z_S) = \langle z_S, \Pi_S z_S \rangle. \quad (19)$$

Equation (18) defines the induced observation geometries of the teacher and student systems. Under PPS, knowledge distillation can therefore be interpreted as approximate preservation of operator-induced accessibility structure between these induced geometries. Importantly, this formulation shifts the focus of distillation from output agreement alone to structural compatibility between observation operators.

4.3 Operator Consistency and Intertwining Relations

When the teacher and student latent spaces differ, we introduce a linear map

$$\Phi : \mathcal{H}_T \rightarrow \mathcal{H}_S. \quad (20)$$

A natural notion of structural compatibility between teacher and student observation geometries is expressed through the approximate intertwining condition introduced previously in Eq. (6). Equation (6) states that projecting before or after the transfer map yields approximately consistent observable structure. From the PPS viewpoint, this condition expresses approximate preservation of observable structure under representation transfer. The corresponding commutative structure is represented by

$$\begin{array}{ccc} \mathcal{H}_T & \xrightarrow{\Pi_T} & \mathcal{H}_T \\ \downarrow \Phi & & \downarrow \Phi \\ \mathcal{H}_S & \xrightarrow{\Pi_S} & \mathcal{H}_S \end{array}$$

Under exact commutativity, observable quotient structures induced by Eq. (4) are preserved under the transfer map. Approximate commutativity therefore provides a geometric notion of observable consistency between latent representation systems [29, 30, 20]. To empirically examine this interpretation, we compare standard KL-based distillation with a PPS-regularized variant explicitly incorporating the operator-consistency objective

$$L_{\text{PPS}} = \|\Phi\Pi_T - \Pi_S\Phi\|_F^2. \quad (21)$$

The experiment is conducted on CIFAR-10 using a ResNet18 teacher model and a lightweight CNN student model. The map defined in Eq. (20) is implemented as a trainable linear projection jointly optimized with the student network. Figure 2 summarizes the comparison between standard KL-based distillation and PPS-regularized distillation. Panel (a) shows the evolution of the commutativity gap induced by Eq. (21), panel (b) reports predictive accuracy, and panel (c) shows the final operator inconsistency after training. The PPS-regularized formulation consistently reduces operator inconsistency between teacher and student systems while maintaining comparable predictive performance [27, 29, 30]. Importantly, this reduction occurs at the level of induced observation geometry rather than solely at the level of output alignment. These results suggest that successful representation transfer may depend not only on agreement between observable outputs, but also on approximate compatibility between the corresponding observation operators. Within PPS, the primary object of analysis therefore becomes the geometry induced by the operators themselves rather than the output distributions alone. Equation (6) provides one possible formalization of this geometric compatibility.

Rather than claiming universality, this experiment serves as a controlled illustration that operator-level geometric consistency can be introduced as an explicit and measurable quantity within standard representation-transfer pipelines. From this perspective, successful representation transfer may depend not only on output agreement, but also on preservation of the induced observable geometry governing representational accessibility.

5 Structural Limits of Output-Based Interpretability

This section analyzes the structural limits of output-based interpretability from the PPS perspective [2, 3, 20]. The central implication is that explanation methods operating on observable outputs cannot, in general, access the full latent representation space [7, 4, 5, 6]. Instead, they inherit the

PPS Experiment: $\Phi\Pi_T \approx \Pi_S\Phi$ Verification (CIFAR-10)

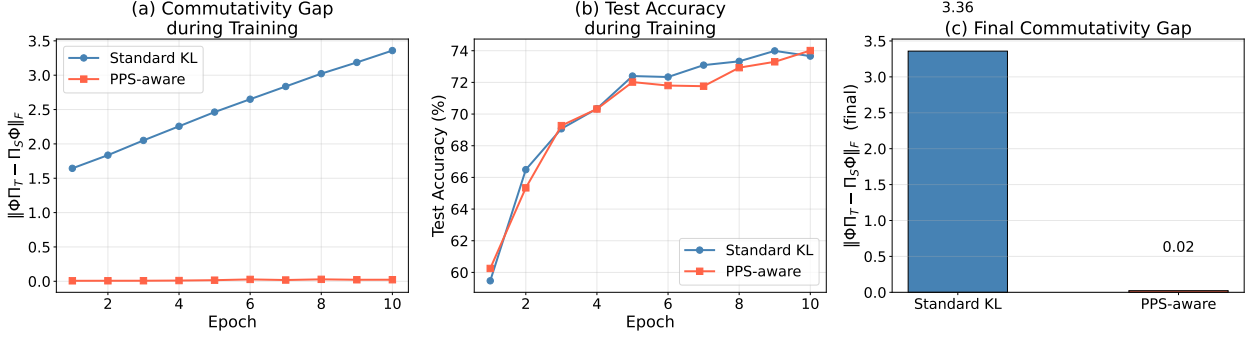


Figure 2: **Operator-consistent knowledge distillation under PPS.** Comparison between standard KL-based distillation and PPS-regularized distillation on CIFAR-10. The PPS formulation explicitly introduces the operator-consistency objective defined in Eq. (21). (a) Evolution of the commutativity gap during training. (b) Test accuracy over epochs. (c) Final operator inconsistency after training. The results indicate that operator-level consistency can be introduced as an explicit optimization objective without degrading predictive performance in this setting.

geometric constraints imposed by the observation operator [20, 14]. Under PPS, observables generated through Eq. (2) depend only on the operator-induced observable geometry [12, 13, 14]. Consequently, latent components contained in $\ker(\Pi)$ are structurally inaccessible to explanation methods derived solely from outputs. This limitation is not a failure of a particular attribution algorithm, but a consequence of projection-mediated observability itself [3, 2].

5.1 Projection-Induced Observability Limits

To make this limitation explicit, we consider the strict projection regime introduced in Eq. (5) [21, 12]. Let $P : \mathcal{H} \rightarrow \mathcal{H}$ be an orthogonal projection operator. For a latent representation $x \in \mathcal{H}$, the projection-induced decomposition is given by

$$x = P(x) + (x - P(x)), \quad (22)$$

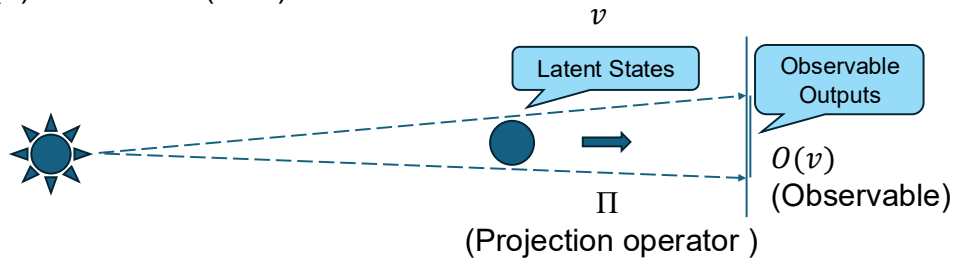
where $P(x)$ is the observable component and $x - P(x)$ lies in the complementary unobservable component. Equation (22) formalizes a key point: interpretability is not an intrinsic property of the latent representation x alone, but of the observation operator acting on it [20, 14]. In this regime, output-based explanations can access only the component preserved by P . Latent directions outside $\text{range}(P)$ do not contribute to observable outputs and therefore cannot be recovered by explanation methods that operate only after projection [7, 4, 5, 6]. This projection-induced decomposition is illustrated conceptually in Fig. 3.

5.2 Observational Equivalence Under Projection

The quotient geometry introduced in Eq. (4) implies that distinct latent states may be observationally indistinguishable [12, 13]. This property can be stated formally as follows.

Proposition 1 (Observational Equivalence Under Projection). *Let the observable functional be defined as in Eq. (2), where Π is a self-adjoint positive semidefinite operator. If $v_1 \sim_{\Pi} v_2$ under*

(1) Definition 1 (PPS).



(2) Definition 2 (Interpretability as Projection-Induced Observation).

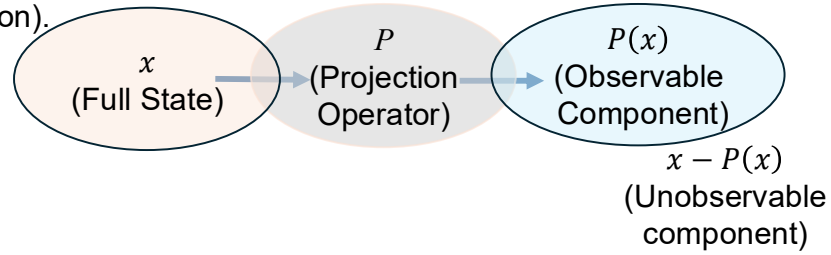


Figure 3: **Conceptual illustration of possible structural implications.** The projection operator P maps a latent representation x to its observable component $P(x)$. The complementary component $x - P(x)$ lies outside the observable subspace and is therefore inaccessible to output-based explanation methods.

Eq. (3), then

$$O(v_1) = O(v_2). \quad (23)$$

Consequently, any explanation method that operates solely on observables generated through Eq. (2) cannot distinguish between latent states belonging to the same equivalence class induced by $\ker(\Pi)$.

Proof. Let $u = v_1 - v_2$ with $u \in \ker(\Pi)$. Then $\Pi u = 0$, and hence $\Pi v_1 = \Pi(v_2 + u) = \Pi v_2$. Substituting this into Eq. (2) yields Eq. (23). Thus, observables cannot distinguish latent states differing only within $\ker(\Pi)$. $\square$

Although algebraically straightforward, Proposition 1 plays a structurally foundational role within PPS because it formalizes observational indistinguishability as a quotient-geometric property induced by the observation operator.

Corollary 1 (Non-Identifiability of Latent Structure). *Under PPS, latent representations are identifiable only up to the quotient geometry defined in Eq. (4). In particular, if $\dim(\ker(\Pi)) > 0$, then the mapping from latent states to observables is non-injective.*

This result shows that output-based interpretability methods cannot reconstruct the full latent representation space [2, 3]. They can only characterize equivalence classes visible through the induced observation geometry.

5.3 Structural Interpretability and Attribution Limits

PPS does not imply that post-hoc attribution methods are invalid. Rather, the framework suggests that such methods are structurally bounded by the induced observation geometry through which observable quantities become accessible. The PPS framework provides a structural explanation for limitations of post-hoc interpretability methods such as SHAP [4], LIME [7], GradCAM [5], and integrated gradients [6]. These methods operate on outputs, activations, or gradients that are already constrained by the observation geometry [2, 3, 20]. Therefore, they characterize only the observable component of the system, not the full latent structure. If a latent factor is encoded in a direction that lies in or near $\ker(\Pi)$, its contribution to observables generated through Eq. (2) is suppressed. As a result, attribution methods may assign negligible importance to a latent component even when that component remains meaningful within the pre-projection latent process [4, 6]. This distinction is important because observational sensitivity and intervention sensitivity need not coincide [39, 36]. A latent component may be unobservable from the output perspective while still affecting downstream behavior through transformations that occur before projection. Thus, absence of attribution should not be interpreted as absence of latent relevance [3, 2]. Figure 4 summarizes this structural limitation of post-hoc explanation.

To illustrate this phenomenon empirically, we consider a controlled synthetic setting in which a latent variable z^* is placed outside the observable subspace. When z^* is excluded from the observable representation, SHAP assigns negligible attribution to it. When the same variable is explicitly included in the observable representation, the attribution becomes non-zero. This change reflects not a change in the underlying latent mechanism, but a change in the accessible observation geometry [20, 14]. Importantly, the intervention in this experiment is applied before projection. Therefore, the experiment does not contradict Proposition 1, which concerns invariance under perturbations confined to $\ker(\Pi)$ after the observation geometry has been fixed. Instead, it demonstrates a separation between latent-space sensitivity and output-based attribution [39, 36].

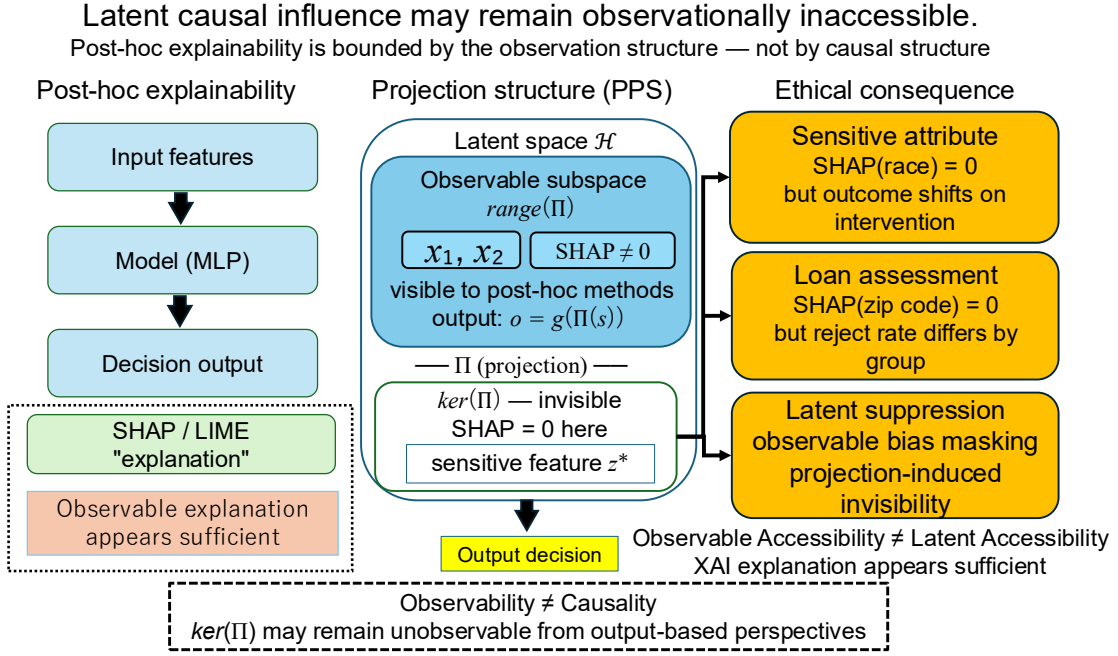


Figure 4: **Structural limitations of post-hoc explainability under PPS.** Post-hoc explanation methods operate only on the observable subspace induced by the observation operator [20]. Latent components weakly represented within this observable geometry may remain inaccessible to output-based attribution methods such as SHAP, LIME, and GradCAM. Consequently, output-based explanations characterize only the observable quotient structure defined in Eq. (4), rather than the full latent representation space.

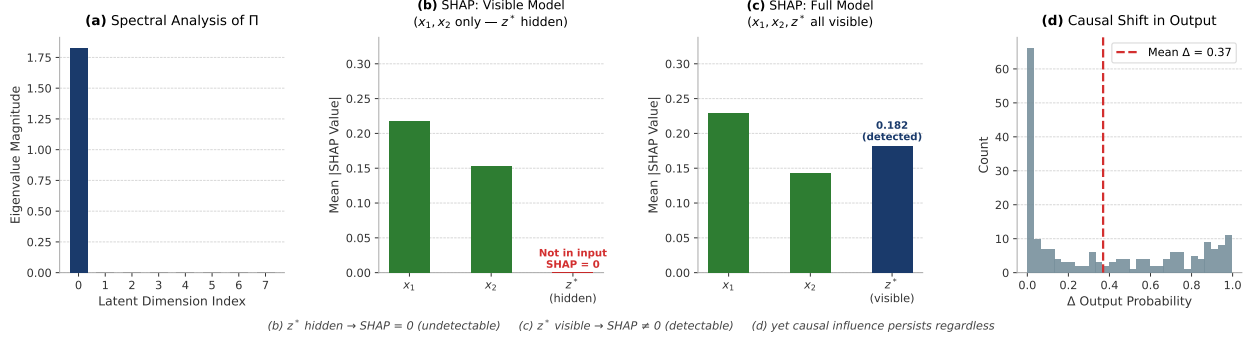


Figure 5: **Empirical illustration of the observability–sensitivity gap under PPS.** (a) Spectral structure of the observation operator showing a non-trivial kernel. (b) When the latent variable lies outside the observable subspace, SHAP attribution becomes negligible. (c) When the same variable is included in the observable representation, attribution becomes non-zero. (d) Latent intervention prior to projection induces measurable output variation, illustrating the distinction between latent-space sensitivity and output-based attribution. Together, these results show that attribution methods inherit structural limitations imposed by projection-mediated observability.

5.4 Toward Structural Accountability

Projection-induced observability limits suggest that output transparency alone is insufficient for characterizing the behavior of representation learning systems [3, 2]. This motivates a shift from output-centric explanation toward structural analysis of the observation operator itself [20, 14]. Within PPS, *structural interpretability* means explicitly characterizing the operator geometry that determines what can become observable [20]. Correspondingly, *structural accountability* concerns whether the observation operator preserves, suppresses, or distorts application-relevant latent directions. This perspective leads to several structural questions:

- Which subspaces are preserved within $\text{range}(\Pi)$?
- Which latent directions are eliminated by $\ker(\Pi)$?
- How does the spectral structure in Eq. (8) regulate representational accessibility?
- How does the effective observable dimension in Eq. (9) constrain explanation and prediction?

The PPS framework therefore suggests that interpretability constraints can be imposed directly at the operator level [33, 34, 35]. Representative examples include spectral constraints controlling eigenvalue decay, subspace constraints enforcing preservation of designated latent directions, and operator regularization penalizing suppression of application-relevant components. Existing approaches can be reinterpreted in this language. Concept Bottleneck Models [33] impose explicit constraints on intermediate observable subspaces, while invariance-based learning methods may be viewed as inducing constraints on observation geometry [35, 36]. PPS provides a unified formulation in which these methods correspond to different ways of shaping the operator that governs observability [20, 14]. From this perspective, interpretability is not merely a property of explanations produced after inference. Rather, it is a property of the observation geometry that determines what can become visible in the first place.

6 Empirical Verification of PPS

These experiments are intentionally designed as controlled operator-level empirical analyses rather than benchmark-oriented evaluations. The objective is not empirical superiority over existing methods, but direct isolation of projection-induced observability effects predicted by PPS. This section presents controlled experiments designed to directly evaluate central predictions of the PPS framework [14, 20]. The experiments focus on three theoretical consequences derived from the operator-theoretic formulation developed in previous sections:

1. invariance of observables under perturbations confined to $\ker(\Pi)$,
2. dependence of observable behavior on the rank and spectral structure of the observation operator,
3. approximate preservation of observable geometry through operator-consistent representation transfer.

Unlike conventional performance-oriented evaluations, the experiments presented here are designed primarily as direct empirical examinations of projection-induced observability predicted by PPS [2, 3].

6.1 Kernel-Invariant Observability

We first evaluate the PPS prediction that observables remain invariant under perturbations confined to the kernel of the observation operator [12, 13]. Consider a latent representation space $\mathcal{H} \cong \mathbb{R}^D$ equipped with a positive semidefinite observation operator $\Pi \succeq 0$ of rank $R < D$. The latent representation is explicitly decomposed into observable and structurally unobservable components:

$$v = v_{\text{obs}} + v_{\text{ker}}, \quad (24)$$

where

$$v_{\text{obs}} \in \text{range}(\Pi), \quad v_{\text{ker}} \in \ker(\Pi).$$

Observables are generated through the PPS quadratic functional defined previously in Eq. (2) [14]. According to Proposition 1, perturbations restricted to $\ker(\Pi)$ should leave observables invariant, whereas perturbations within $\text{range}(\Pi)$ should induce systematic observable variation. To verify this prediction directly, we constructed an 8-dimensional latent space with an observation operator of rank 4. The eigenspectrum therefore consists of four observable modes and four kernel modes. Two perturbation regimes were evaluated:

1. Kernel perturbation

$$v \rightarrow v + \delta_{\text{ker}}, \quad \delta_{\text{ker}} \in \ker(\Pi), \quad (25)$$

2. Observable perturbation

$$v \rightarrow v + \delta_{\text{obs}}, \quad \delta_{\text{obs}} \in \text{range}(\Pi), \quad (26)$$

Figure 6(a) shows the eigenspectrum of the observation operator, separating observable and kernel directions. Figure 6(b) demonstrates that perturbations generated through Eq. (25) leave observables numerically invariant up to machine precision. In contrast, Figure 6(c) shows that perturbations generated through Eq. (26) induce substantial observable variation with approximately

quadratic dependence on perturbation magnitude. This provides a direct empirical visualization of the PPS prediction that latent directions contained in $\ker(\Pi)$ are structurally inaccessible under the induced observation geometry [20, 14]. Finally, Figure 6(d) compares SHAP attribution behavior between restricted and unrestricted models. When the model is explicitly constrained to $\text{range}(\Pi)$, kernel dimensions receive zero attribution by construction. In contrast, unrestricted models distribute attribution across both observable and kernel components. This suggests that attribution methods may reflect representational leakage outside the intended observable subspace structure induced by the operator.

Projection-Induced Observability: Direct Verification of Kernel Invariance (PPS)

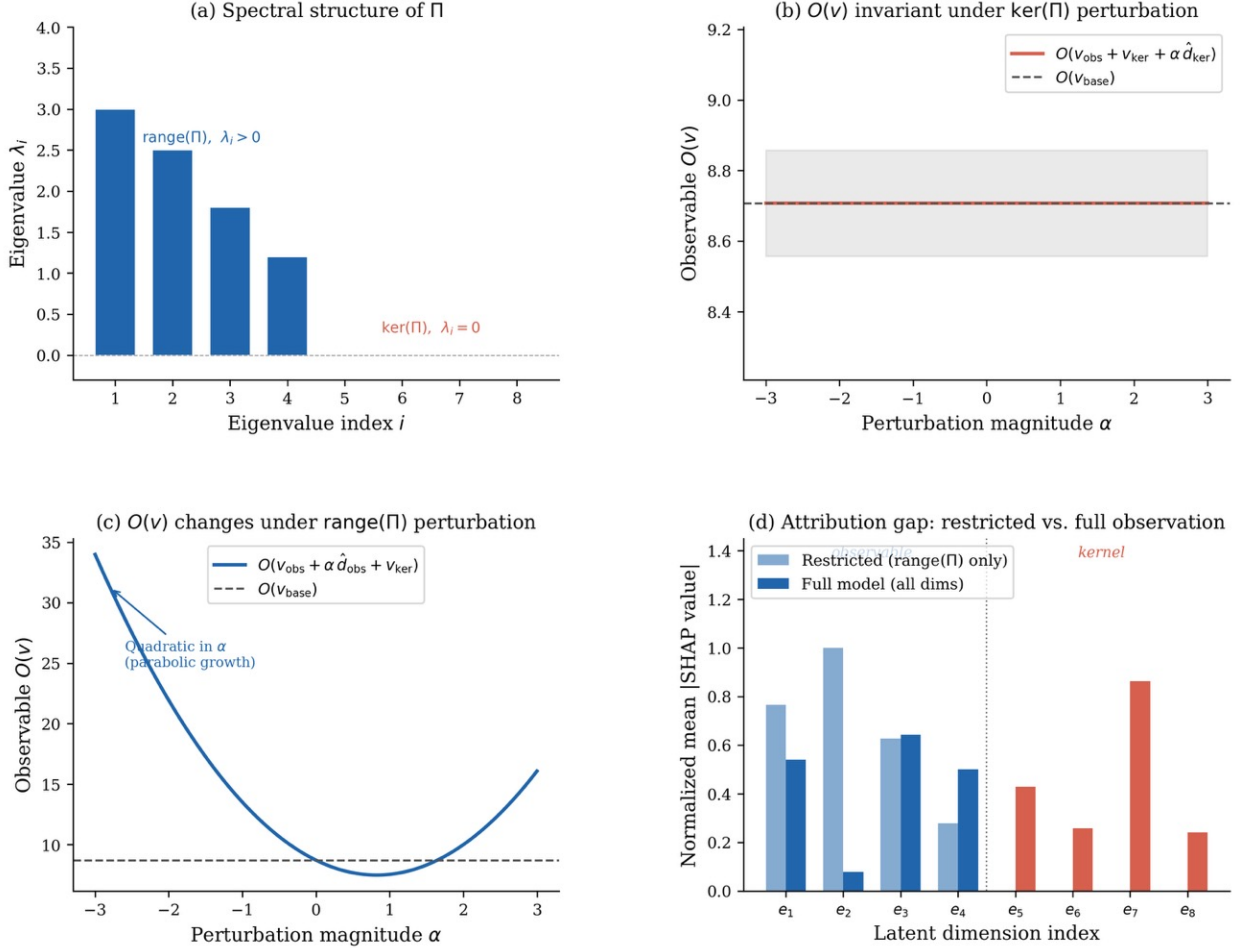


Figure 6: **Direct verification of kernel-invariant observability under PPS.** (a) Eigenvalue spectrum of the observation operator showing decomposition into observable and kernel directions. (b) Observable quantities remain invariant under perturbations confined to $\ker(\Pi)$, directly validating Proposition 1. (c) Perturbations within $\text{range}(\Pi)$ induce systematic observable variation. (d) SHAP attribution comparison between restricted and unrestricted models. The restricted model assigns zero attribution to kernel dimensions, whereas unrestricted models distribute attribution across both observable and kernel components. Together, these results illustrate that observability is governed by operator-induced geometry rather than latent representations alone.

6.2 Rank-Controlled Observable Geometry

We next investigate how the rank structure of the observation operator governs effective observable geometry [18, 17, 38]. Within PPS, the effective observable dimension is determined by Eq. (9). Consequently, observable behavior should depend not on the ambient latent dimensionality itself, but on the rank and spectral structure of the observation operator. To evaluate this prediction, we generated a synthetic classification task in a 16-dimensional latent space where the true signal was confined to an intrinsic subspace of dimension $r^* = 8$. Observation operators Π_r with varying

rank $r \in \{1, \dots, 16\}$ were constructed by progressively truncating the eigenspectrum introduced in Eq. (8). For each rank setting, we evaluated:

1. classification accuracy,
2. attribution leakage into $\ker(\Pi)$,
3. cumulative observable spectral energy.

Figure 7(a) shows that predictive performance increases systematically with $\text{rank}(\Pi)$ and saturates near the intrinsic signal rank $r^* = 8$. Accuracy increases from approximately 0.61 at rank 1 to approximately 0.94 at rank 8, after which additional observable dimensions produce minimal improvement. Figure 7(b) shows the normalized kernel-attribution leakage ratio, defined as the proportion of attribution mass assigned to dimensions outside the observable subspace. Leakage decreases monotonically as $\text{rank}(\Pi)$ increases, indicating that progressively more signal-bearing directions become observable. Figure 7(c) shows the cumulative observable spectral-energy ratio captured by the induced observation geometry. At $r^* = 8$, the observable subspace captures nearly all signal-associated spectral energy, consistent with the observed performance saturation. Finally, Figure 7(d) visualizes cumulative spectral-energy accumulation profiles for multiple rank settings, illustrating how eigenspectrum truncation modifies effective observable geometry. Collectively, these results support the PPS prediction that observable behavior is governed by the spectral structure of the observation operator rather than by latent dimensionality alone [18, 17, 20].

Rank-Controlled Observability: $\text{rank}(\Pi)$ Determines Effective Observable Geometry (PPS)

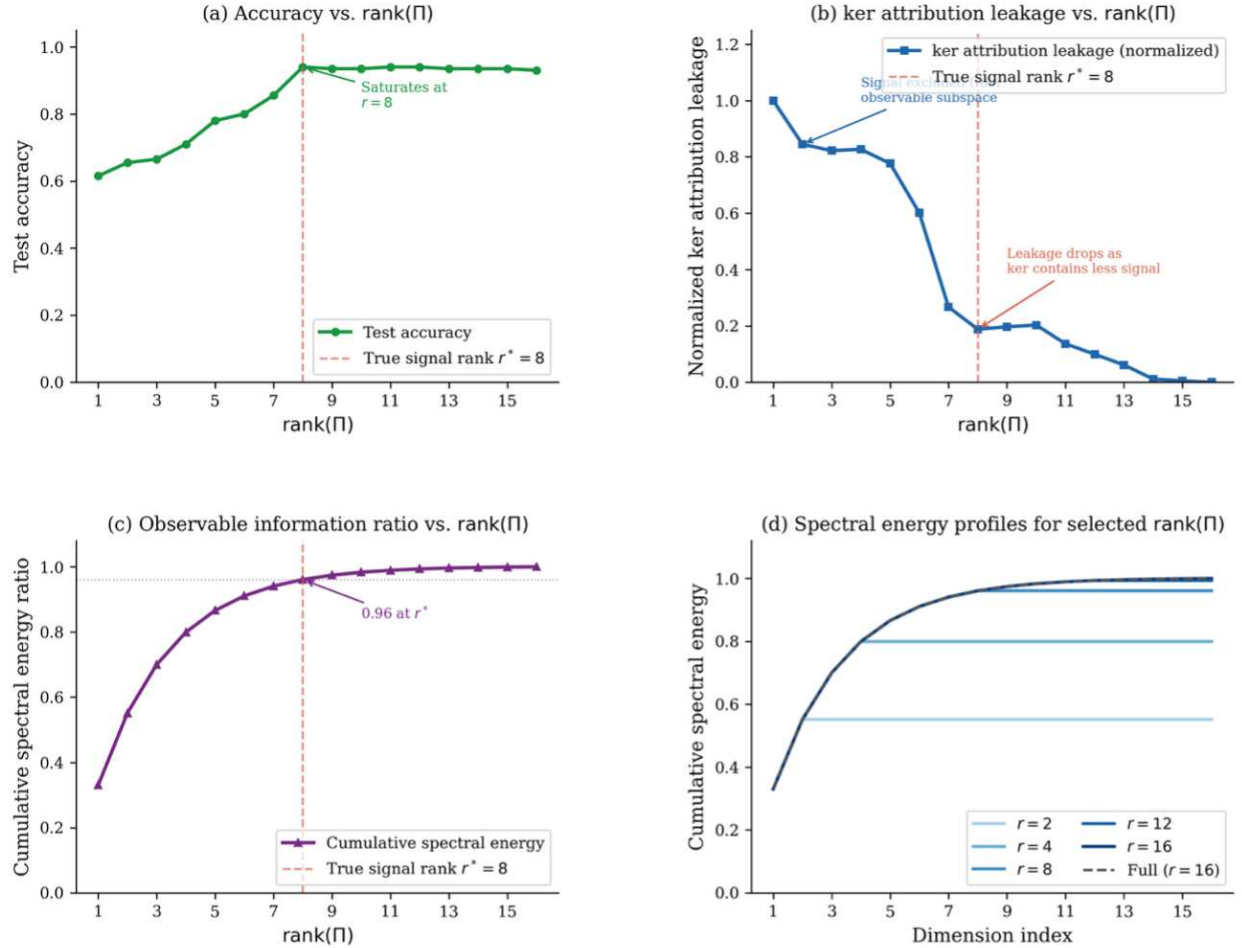


Figure 7: **Rank-controlled observable geometry under PPS.** (a) Classification accuracy as a function of $\text{rank}(\Pi)$. Performance increases systematically with observable dimension and saturates near the intrinsic signal rank. (b) Attribution leakage into $\ker(\Pi)$ decreases as the observable subspace expands. (c) Cumulative observable spectral-energy ratio captured by the induced observation geometry. (d) Spectral-energy accumulation profiles for multiple rank settings, illustrating how eigenspectrum truncation modifies effective observable structure. These results demonstrate that observable behavior is governed by operator-induced geometry rather than ambient latent dimensionality itself.

6.3 Operator-Consistent Knowledge Distillation

Finally, we evaluate the PPS interpretation of representation transfer introduced in Eq. (6) [27, 29, 30]. The experiment compares standard KL-based knowledge distillation with a PPS-regularized formulation incorporating the operator-consistency objective defined in Eq. (21) [27, 28, 29, 30]. A ResNet18 teacher model and a lightweight CNN student model were trained on CIFAR-10. The interspace map introduced in Eq. (20) was implemented as a trainable linear projection jointly optimized with the student network.

Figure 2 summarizes the comparison between standard distillation and PPS-regularized distillation. Panel (a) shows the evolution of the commutativity gap induced by Eq. (21), panel (b) reports predictive accuracy, and panel (c) shows the final operator inconsistency after training. The PPS-regularized formulation consistently reduces operator inconsistency between teacher and student systems while maintaining comparable predictive accuracy [29, 30]. Importantly, the improvement occurs at the level of induced observation geometry rather than solely at the level of output agreement. These results support the PPS interpretation that successful representation transfer may depend not only on observable output matching, but also on approximate compatibility between the corresponding observation operators. Rather than serving as a benchmark-oriented evaluation, the experiment provides a controlled empirical verification that operator-induced geometry can be treated as an explicit and measurable object within representation-transfer pipelines.

7 Discussion

The PPS framework provides an operator-theoretic perspective on observability in representation learning systems [12, 13, 14, 1]. Rather than treating observable outputs as direct reflections of latent representations, PPS models observation as geometry induced through self-adjoint positive semidefinite operators acting on latent Hilbert spaces [12, 14, 20]. This viewpoint connects several previously separate research directions, including information bottleneck theory, observability theory, representation learning, and interpretability analysis [8, 37, 1, 2, 3].

7.1 Relation to Information Bottleneck and Representation Learning

PPS is conceptually related to the information bottleneck (IB) principle [32, 8, 10, 11], which studies compressed representations through mutual-information constraints. Both frameworks emphasize that learned systems preserve only selected aspects of latent information. However, the two approaches differ fundamentally in formulation. Information bottleneck methods characterize compression probabilistically through random variables and mutual-information objectives, whereas PPS characterizes observability geometrically through the spectral structure of observation operators [32, 8, 20, 14]. Within PPS, information accessibility is governed by the quotient geometry introduced in Eq. (4) and the eigenspectrum defined in Eq. (8). Observable information therefore becomes a property of induced operator geometry rather than solely of probabilistic compression [20, 14]. This perspective also relates naturally to modern representation learning. Many learning systems implicitly construct observable subspaces through learned readout operators such as Eq. (16) [1, 15]. From the PPS viewpoint, latent representations should therefore be interpreted together with the observation geometry through which they become accessible.

7.2 Relation to Observability Theory

PPS also exhibits structural parallels with classical observability theory in control systems [37]. Classical observability theory studies whether internal dynamical states can be reconstructed from observable outputs over time. By contrast, PPS focuses on structural accessibility induced by observation operators acting on latent representation spaces [12, 13, 14]. Despite this distinction, both frameworks share a common conceptual principle: observable behavior depends not only on latent states themselves, but also on the structure governing access to those states. Within PPS, this structure is encoded through the operator geometry induced by Π and the quotient structure defined in Eq. (4) [12, 14].

7.3 Interpretability and Structural Accessibility

A central implication of PPS concerns the structural limits of output-based interpretability [2, 3]. Since observables generated through Eq. (2) are invariant under perturbations in $\ker(\Pi)$, they depend only on the quotient structure defined in Eq. (4). The experiments presented in Sections 5 and 6 directly illustrate this phenomenon. Kernel-confined perturbations remain observationally invisible despite latent modification, while eigenspectrum truncation systematically alters attribution behavior and effective observable dimensionality [18, 20]. These results suggest that attribution maps and output-based explanations should not be interpreted as direct reconstructions of latent structure [4, 5, 6, 3]. Rather, they characterize only the observable geometry induced through the observation operator. From this viewpoint, interpretability limitations are not merely algorithmic deficiencies, but structural consequences of projection-mediated observability [2, 3, 20].

7.4 Relation to the Platonic Representation Hypothesis

Recent discussions surrounding the Platonic Representation Hypothesis [40] suggest that independently trained neural systems may converge toward structurally similar latent representations. PPS provides a complementary perspective on this idea. Rather than focusing solely on latent representation similarity, PPS emphasizes the geometry induced by observation operators acting on those representations. From this viewpoint, similarity between systems may arise not only from alignment of latent states themselves, but also from compatibility between induced observable geometries. The operator-consistency relation introduced in Eq. (6) provides one possible formalization of this perspective. The knowledge-distillation experiments further suggest that approximate preservation of operator geometry may serve as a measurable notion of structural compatibility between representation systems [27, 29, 30, 31]. More broadly, PPS suggests that observability may provide a useful geometric language for relating representation learning, interpretability, information accessibility, and projection-mediated inference across diverse systems [20, 1, 14]. In this context, the term “Platonic” is used structurally to emphasize the distinction between latent structure and observable projection, rather than to invoke a metaphysical claim.

7.5 Limitations and Future Work

Several limitations of the present work should be emphasized. First, the current PPS formulation primarily analyzes static observation geometry. Extensions to temporally evolving operators and sequential systems remain future work. Second, the empirical experiments are intentionally controlled and illustrative rather than benchmark-oriented [2, 3]. The primary goal is conceptual validation of projection-mediated observability rather than state-of-the-art predictive performance. Third, the present framework focuses mainly on self-adjoint positive semidefinite operators [14, 13]. More general operator classes, including nonlinear, stochastic, and non-self-adjoint observation mechanisms, remain open directions for investigation. Several future directions appear particularly promising:

- operator-constrained representation learning,
- spectral regularization for interpretable latent geometry,
- projection-aware explanation methods,
- geometric formulations of transfer learning,
- extensions to sequential and multi-modal systems.

In addition, the present formulation primarily focuses on linear self-adjoint observation structures. Extensions to nonlinear operators, stochastic observation processes, non-Hilbert latent geometries, and temporally evolving observability remain important directions for future work.

8 Conclusions

This paper introduced Platonic Projection Structures (PPS), an operator-theoretic framework for analyzing projection-mediated observability in representation learning systems [12, 13, 14]. Within PPS, observable quantities are generated not through direct access to latent representations, but through geometry induced by self-adjoint positive semidefinite observation operators [14, 20]. The framework formalizes this perspective through the observable functional defined in Eq. (2), the quotient structure introduced in Eq. (4), and the spectral decomposition in Eq. (8). From this viewpoint, observability becomes a structural property of induced operator geometry rather than a property of latent states themselves. The framework further provides a unified abstraction spanning quantum measurement, deep learning inference, representation transfer, and interpretability analysis [21, 22, 1, 27, 2]. Quantum systems and neural systems were shown to admit structurally parallel formulations as projection-mediated observation processes, despite substantial differences in physical interpretation [21, 24, 23]. PPS also provides a geometric interpretation of representation transfer through the operator-consistency relation introduced in Eq. (6). The corresponding experiments demonstrated that operator-level consistency can be incorporated explicitly within knowledge-distillation pipelines while preserving predictive performance [27, 29, 30]. A central implication of PPS concerns the structural limits of output-based interpretability [2, 3]. Because observable quantities depend only on the induced observable quotient geometry, attribution methods operating exclusively on outputs necessarily inherit constraints imposed by the observation operator itself. The controlled experiments presented in this work verified several central predictions of the framework, including:

1. kernel-invariant observability,
2. rank-controlled observable geometry,
3. operator-consistent representation transfer.

Collectively, these results support the PPS interpretation that observable behavior is governed by operator-induced geometry rather than by latent representations alone [2, 3]. More broadly, PPS suggests that observability may serve as a unifying geometric principle connecting representation learning, interpretability, information accessibility, and projection-mediated inference across diverse systems [20, 1]. Future work will investigate extensions to nonlinear operators, sequential systems, adaptive observation geometry, multi-modal architectures, and projection-aware interpretability methods. Ultimately, the PPS framework proposes a shift in perspective: rather than asking only what latent representations contain, one may instead ask what the induced observation geometry allows a system to observe [20, 14, 12].

References

- [1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

- [2] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [3] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [4] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.
- [7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.
- [8] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.
- [9] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.
- [10] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.
- [11] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2018.
- [12] Paul R. Halmos. *Introduction to Hilbert Space*. AMS Chelsea Publishing, 1957.
- [13] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics I*. Academic Press, 1980.
- [14] Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.
- [15] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, 2002.
- [16] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.
- [17] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2002.
- [18] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- [19] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2001.

- [20] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.
- [21] John von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, 1955.
- [22] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2010.
- [23] Maria Schuld, Alex Bocharov, Krysta Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.
- [24] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.
- [25] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.
- [26] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.
- [27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.
- [29] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [30] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020.
- [31] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017.
- [32] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
- [33] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020.
- [34] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning (ICML)*, 2018.
- [35] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

- [36] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- [37] Rudolf E. Kalman. On the general theory of control systems. *IRE Transactions on Automatic Control*, 4(3):110–110, 1960.
- [38] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Journal of the ACM*, 58(2):1–37, 2011.
- [39] Judea Pearl. *Causality*. Cambridge University Press, 2009.
- [40] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.