The Davis Manifold

Geometry-First Detection with Compositional Error Budgets

Bee Rosa Davis
NASA Mission Systems Engineer
bee_davis@alumni.brown.edu

Abstract

Many safety-critical detection problems have a common structure: an underlying configuration (e.g., antigenic profile, real human identity, physical pose) evolves along identity-preserving paths, while only indirect, noisy observations are available. Contemporary machine-learning systems typically treat these problems as static classification or black-box sequence modeling, conflating Euclidean similarity in a learned embedding with semantic stability, and providing few explicit regimes of validity, abstention rules, or decomposed error budgets.

We introduce *Davis manifolds* and *Davis systems* as a general framework for geometry-first detection in identity-preserving temporal domains. A Davis manifold is a Riemannian state space equipped with (i) path classes $\mathcal{P}(L)$ describing "benign" identity-preserving trajectories of bounded length L, (ii) a bounded geodesic–Euclidean distortion profile $\varepsilon(L)$ along these paths, and (iii) soft/hard configuration margins $(\kappa_{\mathrm{hard}}, \kappa_{\mathrm{soft}})$ that carve out an explicit ambiguity band where the system must abstain. A Davis system couples such a manifold to (iv) monotone path-based features and a scalar statistic S, (v) a finite-variance tail bound (Cantelli inequality) mapping separation in S to misclassification risk, and (vi) a compositional error budget over geometry, feature linkage, calibration, and abstention failures.

Formally, we (1) define Davis manifolds and Davis systems and state core assumptions (A1–A6) together with non-vacuity conditions that ensure configuration margins are not destroyed by distortion; (2) prove an existence theorem showing that contrastive training (InfoNCE) plus smoothness regularization yields Davis manifolds with explicit $\varepsilon(L) \leq K(\lambda)L$ dependence on regularization strength and path horizon; (3) derive finite-variance detection bounds in which posterior correctness decomposes into a Cantelli term and a multiplicative error budget $(1 - E_{\mathrm{geom}})(1 - E_{\mathrm{link}})(1 - \xi)(1 - \zeta)$ corrected by an independence slack δ_{indep}; (4) analyze the trade-off between path length L and distortion $\varepsilon(L)$ and describe how to choose an operational horizon $L_\star$ that balances coverage against geometric stability; and (5) instantiate the framework in two domains—HERALD for viral antigenic drift and VIDAR for deepfake detection via identity trajectories—as worked Davis systems.

Scope: This manuscript is conceptual. It states definitions, assumptions, theorems, and proposed empirical protocols (including distortion audits, error-budget estimation, and abstention monitoring) but does not report retrospective or prospective performance metrics. Within this scope, Davis manifolds provide a unifying geometric foundation for identity-preserving temporal detection with explicit regimes of validity, abstention as a first-class behavior, and compositional error budgets that make failures legible and falsifiable.

1 Introduction

Many high-stakes learning problems involve *identity-preserving temporal processes*: viruses drifting through antigenic space while remaining the "same" lineage; faces moving through pose and illumination while remaining the same person; robots traversing a workspace while preserving object identity. In all of these cases, what matters is not just what an observation *looks like* in isolation, but how it *moves* in a latent space where distance has semantic meaning.

Standard pipelines blur three distinct questions:

1. **Construction**: How do we construct a representation in which distances and paths correspond to meaningful functional changes?

2. **Traversal**: Given such a representation, how do we monitor trajectories and build detection statistics?

3. **Guarantees**: Under what *explicit* assumptions can we bound risk, expose failure modes, and abstain when the regime of validity is violated?

Most deep learning work optimizes traversal—architectures and losses—on top of whatever representation emerges, and then reports empirical metrics. Geometry, if present, is implicit and un-audited; uncertainty, abstention, and error budgets are usually bolted on post hoc.

Key idea. We propose to reverse this order. A *Davis manifold* is a learned (or inherited) Riemannian manifold on which:

- geodesic distance d_g is constructed to reflect semantic change along identity-preserving paths;

- Euclidean computations in an ambient embedding are guaranteed to approximate d_g within a *bounded-distortion regime* parameterized by $\varepsilon(L)$, the distortion radius for paths of length at most L;

- configuration regions ("similar", "changed", "ambiguous") are separated by soft/hard margins $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$, with ambiguity explicitly mapped to abstention;

- downstream detection statistics admit finite-variance risk bounds with decomposed error budgets—geometry, linkage, calibration, and abstention each have their own terms.

A *Davis system* couples such a manifold with path families, features, aggregation, and abstention policies to form an end-to-end detector whose guarantees are conditional, falsifiable, and operational.

HERALD—a framework for geometry-based viral surveillance—and VIDAR—a framework for Riemannian identity dynamics in deepfake detection—are concrete instantiations of this blueprint in two very different domains. In this manuscript we abstract their common structure into a domain-agnostic theory.

1.1 Construction vs. Traversal

We distinguish sharply between:

Construction (training time). Learn a representation f_ϕ and induced metric g_ϕ such that:

1. InfoNCE-style or margin-based objectives separate configurations in distance (Theory I);

2. smoothness regularization on f_ϕ controls curvature and yields explicit distortion bounds $\varepsilon(L)$ on a path family $\mathcal{P}(L)$;
3. soft/hard configuration margins $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$ and an ambiguity band are identified and validated.

Traversal (runtime). Given observations (x_t), we:

1. embed them via f_ϕ and follow paths on M along $\mathcal{P}(L_\star)$, where $L_\star$ is an operational path horizon;
2. extract windowed features that are monotone functions of distance or curvature along these paths;
3. fuse them via probability-integral transforms (PIT) into a scalar statistic S, map S through a calibrated link to risk, and trigger abstention when distortion, separation, or support fall outside validated regimes.

In a Davis system, Euclidean traversal acquires semantic meaning *only because* the geometry was constructed and regularized to make it so, and *only within* a bounded-distortion regime made explicit and auditable.

1.2 Running Examples: HERALD and VIDAR

We ground the theory in two previously proposed systems that follow this construction-first philosophy.

HERALD (viral surveillance). HERALD learns a pullback Riemannian manifold on viral sequences such that geodesic distances capture antigenic relationships under neutralization data, with Jacobian/Laplacian regularization controlling geodesic–Euclidean distortion. A PIT-fused drift statistic $D(t)$ monitors movement relative to vaccine anchors; Cantelli bounds map separation in $D(t)$ to immune-escape and dominance risk with an explicit error budget over drift detection, escape linkage, calibration, and coverage.

VIDAR (deepfake detection). VIDAR treats face-recognition embeddings as points on the hypersphere S^{d-1}, views short clips as identity trajectories on a Riemannian identity manifold, and constructs Riemannian identity manifold (RIM) features (velocity, acceleration, principal-geodesic residuals, subspace leakage) alongside auxiliary detectors. PIT-normalized features are fused into a statistic $T(Z)$, calibrated to a synthesis probability $\hat{p}$, and combined with bounded-distortion and separation assumptions into a compositional error budget for high-risk alerts.

Both HERALD and VIDAR obey the same high-level pattern:

$$\begin{aligned} \text{contrastive} + \text{smoothness} &\Rightarrow \text{Davis manifold with } \varepsilon(L) \\ &\Rightarrow \text{path features and separation } \Delta_S \\ &\Rightarrow \text{finite-variance risk bounds} + \text{error budget.} \end{aligned}$$

The present work abstracts this pattern, yielding a reusable theory for any domain with identity-preserving temporal structure.

1.3 Contributions

Within this unified perspective, the paper makes the following contributions.

- **C1 (Davis manifolds and systems).** We define Davis manifolds as Riemannian manifolds with path-class-specific distortion bounds $\varepsilon(L)$ and soft/hard configuration margins $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$, and Davis systems as detectors that operate on such manifolds with explicit ambiguity and abstention behavior.

- **C2 (Existence via InfoNCE + smoothness).** We prove a two-part existence result: (i) InfoNCE-style contrastive training yields an identity distance gap Δ_{emb}; (ii) smoothness regularization on the embedding's Jacobian/metric yields curvature bounds that translate into explicit distortion functions $\varepsilon(L) \leq C' K(\lambda) L$ along path families $\mathcal{P}(L)$. Together these yield Davis manifolds with tunable trade-offs between expressivity and distortion.

- **C3 (Finite-variance detection guarantees with error budgets).** We derive finite-variance Cantelli bounds that map separation Δ_S in a scalar detection statistic to misclassification risk, decomposing failure into four components: geometry error E_{geom}, linkage error E_{link}, calibration error ξ, and abstention failure ζ, with a correlation slack δ_{indep} capturing departures from independence.

- **C4 (Path-horizon optimization).** We formalize the trade-off between path length L (coverage and detection power) and distortion $\varepsilon(L)$ (geometric stability), and define an operational path horizon $L_\star$ as the maximizer of a lower bound on detection performance subject to a distortion constraint. A worked toy example illustrates how $L_\star$ emerges from this trade-off.

- **C5 (Operational protocols and instantiations).** We provide protocols for estimating each error-budget component, monitoring distortion and separation, and enforcing abstention; we show how HERALD and VIDAR instantiate all abstract objects via explicit tables mapping symbols to concrete components.

Table 1: Contributions of Davis systems relative to representative prior work.

Aspect	**Typical prior work**	**Davis manifolds / systems**
Geometry	Embeddings with implicit geometry; Euclidean metrics used for convenience	Explicit Riemannian geometry with path-class distortion bounds $\varepsilon(L)$ and validated regimes of validity
Temporal structure	Sequence models (RNNs, transformers) without geometric guarantees	Manifold-valued paths with soft/hard configuration margins and identity-preserving path families $\mathcal{P}(L)$
Separation	Empirical margins or AUC/F1; no analytic link to training objective	Margin-to-separation results: InfoNCE $\Rightarrow$ distance gap Δ_{emb} $\Rightarrow$ statistic separation Δ_S
Risk bounds	Concentration inequalities or asymptotics detached from system design	Finite-variance Cantelli bounds integrated into the detector, with explicit non-vacuity conditions
Error budgets	Single aggregate error metric; failure modes implicit	Decomposed $(E_{\text{geom}}, E_{\text{link}}, \xi, \zeta, \delta_{\text{indep}})$ with estimation protocols and abstention thresholds
Abstention	Often omitted or heuristic confidence thresholds	Abstention as a designed outcome tied to distortion, feature range, separation, and error-budget vacuity
Cross-domain framing	Separate methods per domain (e.g., surveillance, forensics)	Single theoretical framework instantiated by HERALD (antigenic drift) and VIDAR (identity dynamics)

Throughout, we emphasize that all guarantees are *conditional*: they hold only for clips/variants/trajectories that lie within empirically validated distortion, separation, and calibration regimes, and are paired with explicit procedures for auditing those regimes.

1.4 Related Work and Historical Context

Metric learning and contrastive objectives. Classical metric learning and contrastive methods learn distances so that similar pairs are close and dissimilar pairs are far, often in Euclidean space with Mahalanobis metrics or angular margins. Davis systems build on this tradition but differ in *what* is guaranteed: instead of stopping at embedding quality, we use contrastive margins as one step in a chain that leads to path-wise distortion bounds, separation in a scalar statistic, and explicit probability guarantees.

Riemannian geometry in statistics and machine learning. Riemannian geometry has long been used to model curved parameter spaces and structured data: from Fisher-information manifolds and information geometry, to learning on SPD matrices, Lie groups, and hyperbolic spaces. Existing Riemannian ML typically focuses on optimization or representation advantages; Davis manifolds instead emphasize *regimes of validity* (via $\varepsilon(L)$), configuration margins, and path-based guarantees for temporal processes.

Temporal ML, sequence modeling, and manifold-valued time series. Recurrent networks, transformers, and dynamical systems models provide powerful sequence models, and there is growing work on manifold-valued time series (e.g., pose trajectories, SPD flows). However, these methods rarely expose explicit distortion regimes, abstention policies, or decomposed error budgets. Davis systems treat temporal evolution as paths on a purpose-built manifold and make the mapping from geometry to detection guarantees explicit.

Uncertainty, selective prediction, and conformal methods. Calibration, selective prediction, and conformal prediction give tools for uncertainty-aware decisions and coverage guarantees. Davis systems integrate these ideas but tie abstention and calibration back to geometric assumptions: when distortion or separation estimates fail, the system is required to abstain, and this behavior appears as a term (ζ) in the error budget.

Domain-specific precursors: HERALD and VIDAR. HERALD and VIDAR are domain-specific blueprints for geometry-first surveillance and forensics, respectively, each combining learned or inherited geometry, PIT-normalized fusion, finite-variance Cantelli bounds, and explicit error budgets. This paper extracts the common theory that underlies both, generalizes it, and provides a framework that can be applied to other domains such as robotics, medical time series, or audio-visual monitoring.

Historical context. Conceptually, Davis manifolds sit at the intersection of several intellectual lineages:

- *Riemannian geometry*, originating with Riemann and Cartan, which formalized curved spaces and geodesics as models for physical and statistical structure;
- *information geometry*, which interprets statistical models as Riemannian manifolds endowed with Fisher metrics;
- *metric and representation learning*, which learn distances and embeddings tailored to tasks;
- *safety and reliability in ML*, including conformal prediction, selective classification, and robustness analysis.

Davis manifolds contribute a missing piece: a unified, path-based framework that links representation construction, distortion control, temporal dynamics, and compositional error budgets, with abstention and vacuity conditions treated as first-class objects.

1.5 When to Use Davis Systems vs. Standard ML

Davis systems are not a universal replacement for standard ML pipelines. They are most useful when geometry, temporal structure, and explicit guarantees matter more than raw accuracy on a static benchmark.

Table 2: When standard ML suffices vs. when a Davis system is appropriate.

Scenario	**Standard ML is usually enough when...**	**Davis systems are appropriate when...**
Temporal structure	Inputs are i.i.d. or orderless; no notion of identity-preserving trajectories	Identity or functional state persists over time and trajectories carry semantic information
Stakes and guarantees	Errors are low-stakes; empirical test accuracy is sufficient	Decisions are safety-critical and require auditable assumptions, abstention, and lower bounds on correctness
Geometry	Distances are proxies for similarity only empirically	You need distances and paths that *mean* something (e.g., antigenic drift, identity change) with a regime where Euclidean approximations are certified
Multi-signal fusion	A single dominant signal drives performance; simple ensembling works	Multiple heterogeneous detectors must be fused with principled normalization (PIT), linkage guarantees, and interpretable error budgets
Uncertainty and abstention	It is acceptable to always output a label, even when uncertain	The system must abstain when assumptions fail (distortion high, features OOD, separation collapses), and coverage must be tracked explicitly
Cross-domain transfer	Model is deployed in a narrow, fixed context	You anticipate re-using the geometric construction across domains (e.g., between pathogens, media types, sensors) with revalidated assumptions

Concrete examples:

- Image classification on static benchmarks, generic recommendation systems, and short-text sentiment analysis typically fall on the “standard ML” side.
- Viral surveillance, deepfake forensics, medical monitoring where identity and physiology evolve smoothly, and robotics with identity-preserving world models are natural candidates for Davis systems.

1.6 Roadmap and Reading Guide

The remainder of the paper is structured in three acts.

- **Act I (Framework).** Section **??** defines Davis manifolds and Davis systems, introduces path families $\mathcal{P}(L)$, distortion functions $\varepsilon(L)$, and configuration margins $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$, and provides a parameter table and notation index together with HERALD/VIDAR instantiation tables.
- **Act II (Theory).** Section **??** proves existence and construction results (Theorem **??**) from InfoNCE + smoothness; Section **??** derives finite-variance detection bounds (Theorem **??**); Section **??** analyzes path-horizon optimization and the trade-off between coverage and distortion.
- **Act III (Synthesis).** Section **??** discusses operational protocols for error-budget estimation, distortion audits, and abstention monitoring; examines limitations, failure modes, and the Davis universality conjecture; and outlines open problems and future directions.

Reading guide.

- *Theorists* may focus on Sections **??**–**??** (framework, existence, detection bounds, and path optimization) and the proofs in the appendices.
- *Practitioners* can skim this introduction, then read the definitions and parameter tables in Section **??**, the operational guidelines in Section **??** (error-budget estimation, distortion audits, and abstention policies), and the HERALD/VIDAR instantiation tables in Section **??**.
- *Domain experts* (e.g., virology, forensics, robotics) may start with the running examples in Section **??** and the instantiation tables (Tables **??** and **??**), then consult the theory sections as needed.
- *Reviewers and skeptics* may wish to jump to the limitations and open problems in Section **??** first, particularly the failure modes and Davis universality conjecture discussions.

2 Davis Manifolds and Systems

This section introduces the geometric objects that underlie Davis manifolds and Davis systems. We start with a visual, coordinate-free picture of what the framework is trying to capture, then formalize the observation space, manifold, distances, configuration structure, and path families. We conclude with (i) a compact assumptions box, (ii) parameter ranges grounded by two running examples (HERALD and VIDAR), (iii) a practitioner checklist, and (iv) a notation index.

2.1 Geometric Intuition and Running Examples

At a high level, a Davis manifold is a Riemannian space in which:

- points $z \in \mathcal{M}$ encode *functional states* of a system (e.g., antigenic state of a viral variant; identity state of a face in a video);
- geodesic distance $d_g(z, z')$ reflects a task-relevant notion of change (e.g., immune escape, identity change);
- observed data $x \in \mathcal{X}$ (sequences, frames) are mapped into $\mathcal{M}$ by an encoder ϕ whose geometry is constructed at training time;

Figure 1: **Geometric intuition for Davis manifolds.** *Left:* Curved manifold with geodesic vs. chord distances and bounded-distortion band. *Middle:* Configuration regions with hard and soft margins and an ambiguous band. *Right:* Short benign paths $\mathcal{P}(L)$, some of which cross configuration boundaries and correspond to high-risk events.

- at runtime, we traverse short paths on $\mathcal{M}$ and compute Euclidean or tangent-space statistics that inherit semantic meaning because the underlying geometry was built to make geodesic–Euclidean approximations valid in an audited regime.

For intuition, we imagine a three-panel schematic (Figure **??**):

- *Panel A (Geometry).* A curved 2D surface embedded in $\mathbb{R}^3$ represents $\mathcal{M}$. A geodesic arc connecting z and z' traces the shortest path *on* the surface, with length $d_g(z, z')$. A straight ambient-space chord between the same points has length $\delta(z, z')$. In a bounded-distortion regime, these agree up to a small multiplicative error:

$$(1 - \varepsilon)d_g(z, z') \le \delta(z, z') \le (1 + \varepsilon)d_g(z, z').$$

- *Panel B (Configuration structure).* Regions of $\mathcal{M}$ are colored by a configuration map $c : \mathcal{M} \to \mathcal{C}$ (e.g., antigenic cluster; identity class). A coarser map $h : \mathcal{C} \to \{0, 1, \mathrm{amb}\}$ partitions states into "non-event" (0), "event" (1), and "ambiguous" (amb), with inner and outer radii $\kappa_{\text{hard}}, \kappa_{\text{soft}}$ defining a fuzzy margin: inside κ_{hard} we are confident in label 0 or 1; between κ_{hard} and κ_{soft} the theory recommends abstention.

- *Panel C (Path families).* Short, smooth paths γ of length at most L represent *benign dynamics* (e.g., realistic antigenic drift; natural identity motion). Most paths stay within a single configuration region; a subset cross a margin and correspond to high-risk events. The Davis framework focuses on such path families $\mathcal{P}(L)$ and on how reliably Euclidean/tangent computations along them reflect geodesic behavior.

HERALD and VIDAR will serve as running examples throughout: HERALD operates on a pullback manifold induced by a sequence encoder for viral spikes, with geodesic distance approximating antigenic distance; VIDAR operates on the unit hypersphere $\mathbb{S}^{d-1}$ induced by a face-recognition encoder, with geodesic distance approximating identity difference.

2.2 Basic Objects: Observation Space, Manifold, and Distances

We now formalize the core objects.

Definition 1 (Observation space and manifold)**.** *Let $\mathcal{X}$ denote the observation space (e.g., sequences, frames, clips), and let $\mathcal{M}$ be a d-dimensional Riemannian manifold representing functional states. An encoder $\phi : \mathcal{X} \to \mathcal{M}$ maps observations to states; a coordinate map $\psi : \mathcal{M} \to \mathbb{R}^d$ provides an ambient representation. We write*

$$z = \phi(x) \in \mathcal{M}, \qquad u = \psi(z) \in \mathbb{R}^d.$$

Definition 2 (Metric, geodesic distance, and ambient distance)**.** *Let g denote the Riemannian metric on $\mathcal{M}$ with associated geodesic distance $d_g : \mathcal{M} \times \mathcal{M} \to [0, \infty)$, and define the ambient*

Euclidean distance

$$\delta(z, z') := \|\psi(z) - \psi(z')\|_2.$$

We assume ψ is locally bi-Lipschitz on the regions of interest, so that d_g and δ are comparable along sufficiently short paths.

In many applications, g is a pullback metric induced by a learned embedding $f_\theta : \mathcal{X} \to \mathbb{R}^d$ via $g(x) = J_\theta(x)^\top J_\theta(x)$, where J_θ is the Jacobian. HERALD uses such a construction on sequence space, while VIDAR uses the standard metric on $\mathbb{S}^{d-1}$ induced by ℓ_2-normalization of ArcFace embeddings.

2.3 Configuration Structure and Margins

Davis systems distinguish finely-grained *configurations* from coarser decision labels.

Definition 3 (Configuration map and coarse labels)**.** *Let $\mathcal{C}$ be a (possibly large) configuration set and $c : \mathcal{M} \to \mathcal{C}$ a configuration map. Let*

$$h : \mathcal{C} \to \{0, 1, \mathrm{amb}\}$$

be a coarse labeling map, where 0 *denotes "non-event",* 1 *denotes "event" (e.g., escape, synthetic, anomaly), and* amb *collects borderline or unresolved cases. We write $Y = h(c(z)) \in \{0, 1, \mathrm{amb}\}$ for the induced coarse label.*

In many binary applications, $\mathcal{C}$ refines $\{0, 1\}$ (e.g., multiple antigenic clusters all mapped to "escape" vs. "similar"; many identities mapped to "same" vs. "different"), but keeping $\mathcal{C}$ explicit allows the theory to handle richer configuration structures.

Definition 4 (Hard and soft configuration margins)**.** *Fix reference sets $\mathcal{R}_0, \mathcal{R}_1 \subset \mathcal{M}$ that summarize canonical non-event and event configurations (e.g., vaccine anchors; reference identities). Define the geodesic distance to each set,*

$$d_g(z, \mathcal{R}_y) := \inf_{z' \in \mathcal{R}_y} d_g(z, z'), \quad y \in \{0, 1\}.$$

For radii $0 < \kappa_{\text{hard}} \le \kappa_{\text{soft}}$, the hard and soft margins are:

$$\begin{aligned} \textit{hard region for } y : \quad & d_g(z, \mathcal{R}_y) \le \kappa_{\text{hard}}, \\ \textit{ambiguous band} : \quad & \kappa_{\text{hard}} < d_g(z, \mathcal{R}_y) < \kappa_{\text{soft}} \textit{ for some } y, \\ \textit{far-from-y region} : \quad & d_g(z, \mathcal{R}_y) \ge \kappa_{\text{soft}}. \end{aligned}$$

We require that $h(c(z)) = \mathrm{amb}$ whenever z lies in the ambiguous band for any relevant configuration; in such cases the Davis system is expected to abstain.

A simple illustrative specialization is:

- HERALD: $\mathcal{R}_0$ are vaccine-proximal antigenic states; $\mathcal{R}_1$ are known escape states. The ambiguous band corresponds to partial escape or uncertain assay regimes.
- VIDAR: $\mathcal{R}_0$ are identity-consistent trajectories for a given person; $\mathcal{R}_1$ correspond to clearly mismatched identities. The ambiguous band captures borderline identity similarity under poor lighting or occlusion.

2.4 Benign Path Families and Distortion

Davis manifolds are defined locally along *benign* path families that reflect typical dynamics under the data-generating process.

Definition 5 (Path class and length)**.** *A (piecewise) smooth path is a map* $\gamma : [0, 1] \to \mathcal{M}$*. Its geodesic length is*

$$\ell_g(\gamma) := \int_0^1 \|\dot{\gamma}(t)\|_g \, dt.$$

For $L > 0$*, a benign path class* $\mathcal{P}(L)$ *is a collection of paths such that* $\ell_g(\gamma) \leq L$ *for all* $\gamma \in \mathcal{P}(L)$ *and such that each path is consistent with the domain's notion of "identity-preserving" evolution.*

Examples:

- HERALD: $\mathcal{P}(L)$ are mutation paths with at most L amino-acid substitutions restricted to predefined epitope positions.
- VIDAR: $\mathcal{P}(L)$ are short identity trajectories on $\mathbb{S}^{d-1}$ induced by contiguous frame windows with smooth head motion and stable capture conditions.

Definition 6 (Pathwise distortion and distortion radius)**.** *For a path* $\gamma \in \mathcal{P}(L)$ *and times* $0 \leq s < t \leq 1$*, define the distortion ratio*

$$\rho_\gamma(s, t) := \frac{\delta(\gamma(s), \gamma(t))}{d_g(\gamma(s), \gamma(t))} \quad \textit{whenever } d_g(\gamma(s), \gamma(t)) > 0.$$

The distortion radius *at horizon* L *is*

$$\varepsilon(L) := \sup_{\gamma \in \mathcal{P}(L)} \sup_{0 \leq s < t \leq 1} \big|\rho_\gamma(s, t) - 1\big|.$$

A bounded-distortion regime is one in which $\varepsilon(L)$ is small for the horizon of interest $L = L_\star$. In practice, $\varepsilon(L)$ is not computed exactly but is audited empirically via distortion ratios on validation paths.

Definition 7 (Non-vacuity condition)**.** *Let* $R > 0$ *bound the geodesic radius of interest around reference configurations (e.g., the maximum distance at which the system attempts to operate). A non-vacuity condition for a chosen horizon* $L_\star$ *is*

$$\kappa_{\text{soft}} - 2\varepsilon(L_\star)R > 0,$$

ensuring that distortion does not completely eat the configuration margin.

Section **??** (Theorem **??**) will relate $\varepsilon(L)$ to curvature bounds $K(\lambda)$ induced by smoothness regularization: roughly, $\varepsilon(L) \lesssim C'K(\lambda)L$ for short paths, where λ controls the strength of the regularizer.

2.5 Davis Systems

A Davis system couples a Davis manifold with a detection pipeline and an error budget.

Definition 8 (Davis manifold)**.** *A* Davis manifold *is a tuple*

$$\mathcal{D}_{geom} := (\mathcal{X}, \mathcal{M}, g, \phi, \psi, \mathcal{P}, \varepsilon(\cdot), \mathcal{C}, c, h, \kappa_{\text{hard}}, \kappa_{\text{soft}})$$

consisting of:

(i) observation space $\mathcal{X}$ and manifold $(\mathcal{M}, g)$;

(ii) encoder $\phi : \mathcal{X} \to \mathcal{M}$ and coordinates $\psi : \mathcal{M} \to \mathbb{R}^d$;

(iii) benign path families $\mathcal{P}(L)$ and distortion radius $\varepsilon(L)$ for each L;

(iv) configuration structure $(\mathcal{C}, c, h)$ with hard/soft margins $\kappa_{\text{hard}} \leq \kappa_{\text{soft}}$.

Definition 9 (Davis system)**.** *A* Davis system *augments a Davis manifold with a statistic, decision rule, and error budget:*

$$\mathcal{D} := (\mathcal{D}_{geom}, S, A, E_{geom}, E_{link}, \xi, \zeta, \delta_{indep}, L_\star, \varepsilon_\star, \tau_{vac}),$$

where:

(i) S is a scalar detection statistic derived from features of paths in $\mathcal{P}(L_\star)$ and auxiliary detectors;

(ii) A is a decision rule mapping $(S$, context$)$ to labels in $\{$`info`, `warn`, `high_risk`, `insufficient_signal`$\}$;

(iii) E_{geom} is the probability that geometric assumptions (bounded distortion, smoothness, path validity) fail in ways that materially affect S;

(iv) E_{link} is the probability that S fails to reflect geodesic distance even when geometry is good (feature linkage failure);

(v) ξ quantifies calibration error in the mapping from S to probabilities;

(vi) ζ quantifies abstention failure (probability of not abstaining when assumptions are violated);

(vii) δ_{indep} bounds correlations between these error events;

(viii) $L_\star$ is the operational path horizon and $\varepsilon_\star := \varepsilon(L_\star)$;

(ix) τ_{vac} is a vacuity threshold beyond which the compositional bound is declared non-informative.

Theorems in Sections **??**–**??** will connect (i) contrastive training and smoothness to $(\Delta_g, \varepsilon_\star, \kappa_{\text{hard}}, \kappa_{\text{soft}})$ (Theory I), (ii) separation in S to finite-variance misclassification bounds (Theory II), and (iii) these quantities to an end-to-end correctness bound for high-risk alerts under the error budget $(E_{\text{geom}}, E_{\text{link}}, \xi, \zeta, \delta_{\text{indep}})$ (Theory III).

2.6 Core Assumptions (Summary Box)

For reference, we collect the core assumptions used throughout the paper.

Symbol	Meaning	HERALD (illustrative)	VIDAR (illustrative)
$L_\star$	Path horizon	5–10 AA subs	0.2–0.5 rad
$\varepsilon_\star$	Distortion at $L_\star$	0.05–0.15	0.05–0.15
κ_{hard}	Hard margin radius	≈ 0.3 (antigenic units)	≈ 0.1 rad
κ_{soft}	Soft margin radius	≈ 0.5 (antigenic units)	≈ 0.2 rad
α	Feature-slope aggregate	2–5	3–8
E_{geom}	Geometry error	< 0.10	< 0.15
E_{link}	Linkage error	< 0.10	< 0.10
ξ	Calibration error	< 0.05	< 0.05
ζ	Abstention failure	< 0.05	< 0.05
δ_{indep}	Independence slack	≈ 0.02	≈ 0.03
τ_{vac}	Vacuity threshold	0.7	0.7

Table 3: **Representative parameter scales** for two Davis systems (HERALD and VIDAR). Values are illustrative and context-dependent; they are included to provide intuition for typical orders of magnitude, not as reported estimates.

Core Assumptions for Davis Systems

- **A1 (Finite variance).** The detection statistic S and relevant distances satisfy $\mathrm{Var}(S \mid Y = y) < \infty$ and $\mathrm{Var}(\delta \mid Y = y) < \infty$ for $y \in \{0, 1\}$.
- **A2 (Smoothness–curvature control).** The encoder induces a metric g whose curvature along benign paths is bounded by $K(\lambda)$, where λ is a smoothness-regularization strength, and hence $\varepsilon(L) \lesssim C'K(\lambda)L$ for short paths.
- **A3 (Feature monotonicity on operational range).** For the features used to build S, there exist slopes $m_k > 0$ such that, on an operational range $[d_{\min}, d_{\max}]$ of geodesic distances, each feature is approximately monotone: $\phi'_k(d) \geq m_k$ for $d \in [d_{\min}, d_{\max}]$.
- **A4 (Bounded label noise).** Coarse labels $Y \in \{0, 1\}$ have symmetric mislabeling at rate $\eta < \eta_{\max} < \frac{1}{2}$; separation bounds incorporate an $O(\eta)$ slack.
- **A5 (Non-vacuity of configuration margins).** For the chosen horizon $L_\star$ and radius R, the non-vacuity condition $\kappa_{\text{soft}} - 2\varepsilon(L_\star)R > 0$ holds.
- **A6 (Approximate independence of error sources).** The error events corresponding to $E_{\text{geom}}, E_{\text{link}}, \xi, \zeta$ are independent up to a total-variation slack δ_{indep}, or else a conservative union bound is used with an explicit vacuity threshold τ_{vac}.

2.7 Parameter Scales and Example Ranges

To anchor the abstract symbols, Table **??** lists representative parameter ranges for two instantiated Davis systems: HERALD (viral antigenic drift) and VIDAR (identity dynamics in video). These values are illustrative, not empirical claims; they serve only to convey typical scales and to suggest orders of magnitude.

Component	HERALD instantiation (viral surveillance)
$\mathcal{X}$	Viral spike protein sequences
$\mathcal{M}$	Pullback manifold induced by a sequence encoder
g	Pullback metric $g = J_\theta^\top J_\theta$
ϕ	Learned encoder from sequences to $\mathbb{R}^d$ (contrastive + proxy)
ψ	Identity map on $\mathbb{R}^d$ (latent coordinates)
$\mathcal{C}$	Antigenic clusters / functional phenotypes
c	Map from z to antigenic configuration (e.g., cluster assignment)
h	Coarse map: similar vs. escape vs. ambiguous
$\mathcal{P}(L)$	Epitope-restricted mutation paths with $\leq L$ substitutions
$L_\star$	Operational substitution horizon (e.g., 5–10 AA changes)
S	Drift statistic $D(t)$ from PIT-fused sequence/antigenic/structural signals
A	Alert rule for dominance risk at horizon T
Error budget	E_{geom}: manifold distortion; E_{link}: escape linkage; ξ: calibration; ζ: abstention failures

Table 4: **HERALD as a Davis system.** Mapping of abstract components to viral surveillance instantiation.

Component	VIDAR instantiation (deepfake forensics)
$\mathcal{X}$	Short talking-head video clips (with optional audio)
$\mathcal{M}$	Identity manifold $\mathbb{S}^{d-1}$ (unit hypersphere)
g	Standard Riemannian metric on $\mathbb{S}^{d-1}$
ϕ	Face-recognition encoder (ArcFace-style, frozen or fine-tuned)
ψ	Identity embedding in $\mathbb{R}^d$ with ℓ_2-normalization
$\mathcal{C}$	Identity and capture-condition configurations
c	Map from z to identity and context configuration
h	Coarse map: same vs. different identity vs. ambiguous
$\mathcal{P}(L)$	Smooth identity trajectories with arc length $\leq L$
$L_\star$	Operational arc-length horizon (e.g., 0.2–0.5 rad)
S	Detection statistic $T(\bar{Z})$ from RIM features + auxiliary detectors
A	Decision rule for $\{\texttt{info}, \texttt{warn}, \texttt{high\_risk}, \texttt{insufficient\_signal}\}$
Error budget	E_{geom}: identity-trajectory geometry; E_{link}: feature linkage; ξ: calibration; ζ: abstention failures

Table 5: **VIDAR as a Davis system.** Mapping of abstract components to deepfake detection instantiation.

2.8 Example Instantiations: HERALD and VIDAR

We now summarize, at a high level, how HERALD and VIDAR instantiate the abstract components of a Davis system. These tables serve as running examples referenced in later sections.

2.9 Practitioner Checklist: Building a Davis System

To make the framework actionable, we provide a high-level checklist for constructing a Davis system in a new domain. Later sections will give theorems and protocols; this box summarizes the workflow.

Practitioner Checklist for Davis Systems

1. **Define observations and identity preservation.** Specify $\mathcal{X}$ and what it means for two observations to share an underlying "identity" over time (e.g., same virus lineage; same person).

2. **Choose/train an encoder.** Select or train $\phi : \mathcal{X} \to \mathbb{R}^d$ with contrastive objectives and smoothness regularization to induce a meaningful geometry.

3. **Define benign path families.** Specify $\mathcal{P}(L)$ as identity-preserving trajectories and choose a candidate path horizon $L_\star$.

4. **Audit distortion.** Empirically estimate distortion ratios along paths in $\mathcal{P}(L)$ on validation data; pick $L_\star$ such that $\varepsilon(L_\star)$ and the non-vacuity condition are acceptable.

5. **Define configurations and margins.** Construct configuration map c and coarse labels h, and estimate operational values for $\kappa_{\text{hard}}, \kappa_{\text{soft}}$.

6. **Design features and statistic.** Build features from paths in $\mathcal{P}(L_\star)$ and auxiliary detectors; define a scalar statistic S (e.g., via PIT-normalized fusion).

7. **Estimate error budget.** On held-out data, estimate or upper-bound $(E_{\text{geom}}, E_{\text{link}}, \xi, \zeta, \delta_{\text{indep}})$ and decide on a vacuity threshold τ_{vac}.

8. **Set thresholds and abstention policies.** Choose decision thresholds for S, distortion gates, and abstention rules consistent with the error budget and application costs.

9. **Monitor and retrain.** In deployment, track distortion, coverage, error-budget estimates, and slice behavior; retrain or narrow scope when core assumptions or non-vacuity conditions degrade.

2.10 Notation Index

For convenience, Table **??** summarizes the main symbols used throughout the paper.

3 Existence: From Contrastive Training and Smoothness to Davis Manifolds

Section **??** introduced Davis manifolds and Davis systems abstractly. In this section we show how they arise from a concrete construction: contrastive training with a smoothness regularizer. Informally, we prove that (i) a margin in a contrastive loss induces separation in the learned geometry, and (ii) a smoothness penalty bounds geodesic–Euclidean distortion along paths of bounded length. Together these yield an existence theorem: under suitable hyperparameters, training produces a Davis manifold in the sense of Definition **??** with a non-vacuous configuration margin (Definition **??**) and a controlled path family (Definition **??**).

3.1 Setup: Embedding, In-Distribution Region, and Path Families

We consider a parametric embedding

$$f_\theta : \mathcal{X} \to \mathbb{R}^d$$

Symbol	Meaning
$\mathcal{X}$	Observation space (sequences, frames, clips)
$\mathcal{M}$	Riemannian manifold of functional states
$\phi : \mathcal{X} \to \mathcal{M}$	Encoder from observations to manifold
$\psi : \mathcal{M} \to \mathbb{R}^d$	Coordinate map / ambient representation
g	Riemannian metric on $\mathcal{M}$
$d_g(z, z')$	Geodesic distance induced by g
$\delta(z, z')$	Euclidean distance in ambient space $\mathbb{R}^d$
$\mathcal{P}(L)$	Benign path family with geodesic length $\leq L$
$\varepsilon(L)$	Distortion radius at path horizon L
$\mathcal{C}$	Configuration set (fine-grained states)
$c : \mathcal{M} \to \mathcal{C}$	Configuration map
$h : \mathcal{C} \to \{0, 1, \text{amb}\}$	Coarse label map (non-event, event, ambiguous)
$\kappa_{\text{hard}}, \kappa_{\text{soft}}$	Hard and soft configuration margins
Y	Coarse label $Y = h(c(z)) \in \{0, 1, \text{amb}\}$
S	Scalar detection statistic derived from features and auxiliaries
Δ_S	Class separation in S (difference in class means)
E_{geom}	Probability of geometric failure affecting S
E_{link}	Probability of feature-linkage failure
ξ	Calibration error in mapping S to probabilities
ζ	Abstention failure probability
δ_{indep}	Independence slack between error sources
$L_\star$	Operational path horizon
$\varepsilon_\star$	$\varepsilon(L_\star)$, distortion bound at horizon $L_\star$
τ_{vac}	Vacuity threshold for total error budget

Table 6: **Notation index** for Davis manifolds and systems.

with parameter vector $\theta \in \Theta$. The image $\mathcal{M}_\theta = f_\theta(\mathcal{X})$ is equipped with the pullback metric g_θ induced by the ambient Euclidean metric, and geodesic distance d_{g_θ} on $\mathcal{M}_\theta$. We write

$$\delta_\theta(x, x') := \left\| f_\theta(x) - f_\theta(x') \right\|_2$$

for the induced Euclidean distance between embedded observations.

In-distribution region. The theoretical guarantees in this section are intended to hold only on an *in-distribution region* $\Omega_{\text{in}} \subset \mathcal{X}$, defined operationally by quality filters and out-of-distribution (OOD) detectors (Section **??**). Concretely, Ω_{in} is the subset of $\mathcal{X}$ on which: (i) we have sufficient training coverage to estimate null distributions and calibration maps; and (ii) runtime distortion audits and quality checks are expected to pass with high probability. All statements below are restricted to $x, x' \in \Omega_{\text{in}}$ unless otherwise noted.

Configuration structure and path families. We assume a configuration map $c : \mathcal{M}_\theta \to \mathcal{C}$ together with a coarse map $h : \mathcal{C} \to \{0, 1, \text{amb}\}$ as in Definition **??**, and a family of paths $\mathcal{P}(L)$ as in Definition **??**, consisting of curves $\gamma : [0, 1] \to \mathcal{M}_\theta$ of geodesic length at most L. In applications, $\mathcal{P}(L)$ encodes "benign" transformations (e.g., short mutation paths, smooth identity trajectories) that preserve configuration except at explicit transition points.

Our goal in this section is to show that for suitable training hyperparameters $(\lambda, L_\star)$ there exist:

- a distortion profile $\varepsilon(L)$ with $\varepsilon(L_\star)$ small enough that Euclidean distances approximate geodesic distances along $\mathcal{P}(L_\star)$, and
- hard/soft configuration margins $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$ that are not eaten by distortion,

so that $(\mathcal{M}_\theta, g_\theta, \mathcal{P}(L_\star), \varepsilon, c, h)$ satisfies Definition **??** and the non-vacuity condition

$$\kappa_{\text{soft}} - 2R\,\varepsilon(L_\star) \;>\; 0, \tag{1}$$

where R is a radius controlling the local neighborhood in which we operate (e.g., a bound on d_{g_θ} between reference configurations and points of interest).

Training objective. We assume θ is obtained by minimizing a composite loss

$$\mathcal{L}(\theta) \;=\; \alpha\,\mathcal{L}_{\text{base}}(\theta) + \beta\,\mathcal{L}_{\text{NCE}}(\theta) + \lambda\,\mathcal{L}_{\text{smooth}}(\theta), \tag{2}$$

where:

- $\mathcal{L}_{\text{base}}$ is a task-specific loss (e.g., cross-entropy on labels derived from c or h),
- $\mathcal{L}_{\text{NCE}}$ is an InfoNCE-style contrastive loss over *configuration-consistent* pairs (x, x^+) and *configuration-changing* negatives (x, x^-), defined below, and
- $\mathcal{L}_{\text{smooth}}$ is a smoothness regularizer that penalizes rapid variation of f_θ and the induced metric g_θ on Ω_{in}.

We now show how $\mathcal{L}_{\text{NCE}}$ yields a configuration-dependent distance gap (Part A), and how $\mathcal{L}_{\text{smooth}}$ bounds distortion along $\mathcal{P}(L)$ (Part B), before combining them into an existence theorem for Davis manifolds.

3.2 Part A: Contrastive Training and Configuration Distance Gaps

This subsection proves the first component of Theorem **??**: a small InfoNCE loss on configuration-respecting pairs induces a margin in the embedding space that separates "same-configuration" from "different-configuration" points.

Let $\mathcal{D}_{\text{cfg}}$ be a distribution over tuples $(x, x^+, x_1^-, \ldots, x_K^-)$ satisfying:

(A1) $c(f_\theta(x)), c(f_\theta(x^+))$ lie in the same coarse configuration under h (e.g., both mapped to 0),

(A2) each $c(f_\theta(x_k^-))$ lies in a configuration that h maps to a *different* label (e.g., 1), and

(A3) all points fall in the in-distribution region: $x, x^+, x_k^- \in \Omega_{\text{in}}$.

Define a similarity score

$$s_\theta(x, x') \;=\; -\frac{1}{2\tau^2}\,\|f_\theta(x) - f_\theta(x')\|_2^2,$$

with temperature $\tau > 0$, and the InfoNCE objective

$$\mathcal{L}_{\text{NCE}}(\theta) \;=\; \mathbb{E}_{(x,x^+,x_1^-,\ldots,x_K^-)\sim\mathcal{D}_{\text{cfg}}}\left[-\log\frac{\exp\{s_\theta(x,x^+)\}}{\exp\{s_\theta(x,x^+)\} + \sum_{k=1}^{K}\exp\{s_\theta(x,x_k^-)\}}\right]. \tag{3}$$

Lemma 1 (Contrastive margin $\Rightarrow$ configuration separation)**.** *Assume:*

(B1) $s_\theta(x, x')$ *is* L_{eff}*-Lipschitz in* $\delta_\theta(x, x') = \|f_\theta(x) - f_\theta(x')\|_2$ *on the range of distances attained on* Ω_{in}*,*

(B2) the empirical InfoNCE loss $\widehat{\mathcal{L}}_{\text{NCE}}(\theta)$ *is within* η *of the population loss* $\mathcal{L}_{\text{NCE}}(\theta)$ *with high probability, and*

(B3) the negative sampling in $\mathcal{D}_{\text{cfg}}$ *respects the configuration structure, in the sense that each* x_k^- *is drawn from a distribution whose* (c, h)*-labels differ from those of* (x, x^+) *with probability at least* $1 - \rho$ *for some* $\rho < \frac{1}{2}$*.*

Then there exist constants $c, C > 0$ *such that, with high probability,*

$$m_{\text{eff}} \;\geq\; \log K - \widehat{\mathcal{L}}_{\text{NCE}}(\theta) - c\sqrt{\frac{C}{n}} - \eta - O(\rho),$$

where m_{eff} *is an effective margin between expected positive and negative scores under* $\mathcal{D}_{\text{cfg}}$*. Moreover, the corresponding expected* configuration distance *gap*

$$\Delta_{\text{cfg}} := \mathbb{E}\big[\delta_\theta(x, x^-) \,\big|\, h \circ c(f_\theta(x)) \neq h \circ c(f_\theta(x^+))\big] - \mathbb{E}\big[\delta_\theta(x, x^+) \,\big|\, h \circ c(f_\theta(x)) = h \circ c(f_\theta(x^+))\big]$$

satisfies

$$\Delta_{\text{cfg}} \;\geq\; \frac{m_{\text{eff}}}{L_{\text{eff}}}.$$

Proof sketch. The proof follows the standard InfoNCE margin-to-separation argument, adapted to the configuration-aware setting and restricted to Ω_{in}. Using $\log(1 + \sum_k e^{\Delta_k}) \geq \log K + \frac{1}{K}\sum_k \Delta_k$ and $\Delta_k = s_\theta(x, x_k^-) - s_\theta(x, x^+)$, one shows that the population InfoNCE loss lower bounds $\log K - m_{\text{eff}}$, so that $m_{\text{eff}} \geq \log K - \mathcal{L}_{\text{NCE}}(\theta)$. Uniform convergence then gives a deviation term $c\sqrt{C/n} + \eta$. Lipschitz continuity of s_θ in $\delta_\theta(\cdot, \cdot)$ converts a score margin into a distance margin, yielding $\Delta_{\text{cfg}} \geq m_{\text{eff}}/L_{\text{eff}}$ up to the contamination factor ρ from occasional mis-sampled negatives. A full proof with explicit constants appears in Appendix A. $\square$

Intuitively, Lemma **??** says: if contrastive training can reliably distinguish same-configuration from different-configuration pairs, then the embedding must allocate a non-trivial distance gap between them. This will later serve as the raw material for configuration margins $\kappa_{\text{hard}}, \kappa_{\text{soft}}$ in Definition **??**.

3.3 Part B: Smoothness Regularization and Bounded Distortion

We now connect the smoothness term $\mathcal{L}_{\text{smooth}}$ in (**??**) to a distortion profile $\varepsilon(L)$ on the path family $\mathcal{P}(L)$.

Smoothness proxy. On the in-distribution region Ω_{in}, we consider a regularizer of the form

$$\mathcal{L}_{\text{smooth}}(\theta) \;=\; \mathbb{E}_{x \sim \mu_{\text{in}}}\big[\|\nabla_x f_\theta(x)\|_F^2\big] \quad \text{or} \quad \mathcal{L}_{\text{smooth}}(\theta) \;=\; \operatorname{Tr}(F^\top L F), \tag{4}$$

where $F = [f_\theta(x_1), \ldots, f_\theta(x_n)]^\top$ and L is a graph Laplacian over Ω_{in}. In both cases, the effect of increasing λ in (**??**) is to penalize rapid variation of the Jacobian $J_\theta(x) = \partial f_\theta(x)/\partial x$ and thus to control the curvature of the pullback metric $g_\theta = J_\theta^\top J_\theta$ on Ω_{in}.

Proposition 1 (Smoothness $\Rightarrow$ bounded distortion along $\mathcal{P}(L)$)**.** *Fix $\lambda > 0$ in (**??**). Suppose that on Ω_{in}:*

(C1) the Jacobian is uniformly bounded: $\|J_\theta(x)\|_F \le B(\lambda)$ for all $x \in \Omega_{\text{in}}$, and

(C2) the covariant derivative of g_θ along any path in $\mathcal{P}(L)$ is bounded: for all $\gamma \in \mathcal{P}(L)$ with image in Ω_{in},

$$\|\nabla_{\dot\gamma} g_\theta(\gamma(t))\| \;\le\; K(\lambda) \quad \textit{for all } t \in [0,1],$$

where $K(\lambda)$ is non-increasing in λ.

Then there exists a constant $C' > 0$ (depending on the geometry of $\mathcal{M}_\theta$ on Ω_{in}) such that for every path $\gamma \in \mathcal{P}(L)$ with $\gamma([0,1]) \subset \Omega_{\text{in}}$,

$$\left|\ell_{g_\theta}(\gamma) - \ell_{\delta_\theta}(\gamma)\right| \;\le\; C' K(\lambda) L^2,$$

and hence, for any endpoints $z, z' \in \gamma([0,1])$,

$$(1 - \varepsilon(L))\, d_{g_\theta}(z,z') \;\le\; \delta_\theta(z,z') \;\le\; (1 + \varepsilon(L))\, d_{g_\theta}(z,z'), \qquad \varepsilon(L) := C'' K(\lambda) L, \tag{5}$$

for some constant $C'' > 0$. In particular, increasing λ decreases $K(\lambda)$ and hence $\varepsilon(L)$ at fixed path length L, while increasing L worsens $\varepsilon(L)$ linearly.

Proof sketch. The assumptions bound both the norm of the Jacobian (controlling local scaling) and the rate at which the metric g_θ can change along paths in $\mathcal{P}(L)$. Standard Riemannian geometry arguments (e.g., comparison with a constant-curvature model space) then imply that the geodesic length of γ deviates from its chordal length by at most a term proportional to $K(\lambda)L^2$, yielding the first inequality. Since $d_{g_\theta}(z,z')$ is the infimum of $\ell_{g_\theta}(\gamma)$ over all connecting paths, the relative distortion between $d_{g_\theta}(z,z')$ and $\delta_\theta(z,z')$ can be bounded by a constant multiple of $K(\lambda)L$, giving (**??**). The dependence of $K(\lambda)$ on λ follows from the role of $\mathcal{L}_{\text{smooth}}$ as a curvature proxy; a detailed argument and references are provided in Appendix A. □

Intuition. Smoothness regularization makes the learned geometry "flatter" on Ω_{in} by penalizing sharp changes in f_θ and hence in g_θ. When curvature is small, geodesics and Euclidean chords agree up to a controlled error: short paths on a nearly-flat manifold are well-approximated by straight lines in the embedding. Proposition **??** captures this intuition quantitatively: stronger regularization (larger λ) shrinks $K(\lambda)$ and therefore the distortion radius $\varepsilon(L)$, while longer path horizons L inevitably pay a price in distortion.

3.4 Combined Existence Theorem for Davis Manifolds

We now combine Lemma **??** and Proposition **??** to obtain an existence theorem: under suitable hyperparameters, training with (**??**) produces a Davis manifold with a non-vacuous configuration margin.

Theorem 1 (Existence of Davis manifolds under contrastive + smoothness training)**.** *Fix a target path horizon $L_\star > 0$ and radius $R > 0$. Assume:*

*(D1) The contrastive conditions of Lemma **??** hold on Ω_{in}, and the resulting configuration distance gap Δ_{cfg} is strictly positive.*

(D2) The smoothness conditions of Proposition **??** *hold on* Ω_{in} *for some* $\lambda > 0$*, yielding a distortion profile* $\varepsilon(L) = C''K(\lambda)L$ *on* $\mathcal{P}(L)$*.*

(D3) The path family $\mathcal{P}(L_\star)$ *consists of paths whose images lie in* $\{z \in \mathcal{M}_\theta : d_{g_\theta}(z, z_0) \leq R\}$ *for some center(s)* z_0 *associated with reference configurations, so that* d_{g_θ} *between relevant points is at most* $2R$*.*

Define hard and soft configuration margins

$$\kappa_{\text{hard}} := \frac{1}{4}\Delta_{\text{cfg}}, \qquad \kappa_{\text{soft}} := \frac{1}{2}\Delta_{\text{cfg}}.$$

Then there exists $\lambda_\star$ *(and hence an associated distortion bound* $\varepsilon_\star := \varepsilon(L_\star)$*) such that for all* $\lambda \geq \lambda_\star$*:*

(E1) The tuple $(\mathcal{M}_\theta, g_\theta, \mathcal{P}(L_\star), \varepsilon, c, h)$ *is a Davis manifold in the sense of Definition* **??** *and Definition* **??***.*

(E2) The non-vacuity condition

$$\kappa_{\text{soft}} - 2R\,\varepsilon(L_\star) \;>\; 0$$

holds, so that configuration changes cannot be entirely eaten by distortion within the operational horizon $L_\star$*.*

Proof sketch. By Lemma **??**, a small InfoNCE loss yields $\Delta_{\text{cfg}} > 0$, so that embeddings of different configurations are, on average, at least Δ_{cfg} farther apart than same-configuration pairs. Choosing $\kappa_{\text{hard}} = \frac{1}{4}\Delta_{\text{cfg}}$ and $\kappa_{\text{soft}} = \frac{1}{2}\Delta_{\text{cfg}}$ ensures that: (i) points within κ_{hard} of a reference configuration z^0 are unlikely to belong to a different coarse configuration, and (ii) points at distance at least κ_{soft} from all same-configuration references are strongly suggestive of a change under h, matching Definition **??**.

By Proposition **??**, for any fixed $L_\star$ and radius R we can increase λ until $K(\lambda)$, and hence $\varepsilon(L_\star)$, is small enough that

$$\kappa_{\text{soft}} - 2R\,\varepsilon(L_\star) \;=\; \frac{1}{2}\Delta_{\text{cfg}} - 2R\,\varepsilon(L_\star) \;>\; 0.$$

This guarantees that even worst-case distortion along paths of length at most $L_\star$ within the radius-R region cannot erase the soft margin. The bounded-distortion inequality in (**??**) and the path-family properties in Definition **??** then verify the remaining conditions of Definition **??**. A complete proof with explicit parameter dependencies appears in Appendix A. □

Theorem **??** formalizes the informal picture from Section **??**: contrastive training carves out configuration separation in the embedding, smoothness regularization keeps the geometry flat enough on $\mathcal{P}(L_\star)$, and a non-vacuity condition ensures that configuration margins survive bounded distortion.

3.5 Remarks and Instantiations

Trade-offs between λ and $L_\star$. The distortion profile $\varepsilon(L) = C''K(\lambda)L$ makes explicit the core trade-off: for a fixed path horizon $L_\star$, increasing λ decreases $K(\lambda)$ and hence $\varepsilon(L_\star)$, tightening the Davis-manifold regime but potentially reducing representation flexibility; for fixed λ, increasing $L_\star$ improves coverage of long-range transformations but worsens distortion. Section **??** returns to this trade-off and formulates an optimization problem for choosing $L_\star$ to maximize downstream detection guarantees (Section **??**) subject to a target distortion budget.

Learned vs. inherited manifolds. Theorem **??** applies most directly when both separation and smoothness are learned (e.g., via InfoNCE and a Jacobian/Laplacian regularizer). In some applications, however, part of the geometry is inherited from a frozen backbone (e.g., a face-recognition or protein-language model). In that case, Lemma **??** may be replaced by an empirical assumption about Δ_{cfg} (as in the inherited-margin regime of HERALD and VIDAR), and Proposition **??** is applied only to fine-tuning layers or to the restricted region Ω_{in} where distortion audits empirically validate a small $\varepsilon(L_\star)$.

Connection to downstream detection guarantees. The Davis manifold produced by Theorem **??** is a *geometric* object: it says nothing yet about detection statistics or risk bounds. Section **??** will introduce features extracted along paths in $\mathcal{P}(L_\star)$, show how a monotone feature mapping propagates configuration separation to separation in a scalar statistic S, and apply finite-variance inequalities to obtain misclassification-risk bounds. The existence theorem here ensures that those features are computed in a regime where Euclidean operations have a meaningful Riemannian interpretation.

HERALD and VIDAR as concrete instances. HERALD and VIDAR provide two domain-specific instantiations of Theorem **??**:

- In HERALD, f_θ embeds viral sequences into a latent space with a pullback metric regularized via Jacobian and Laplacian terms; $\mathcal{P}(L)$ consists of epitope-restricted mutation paths of length at most L; and contrastive pairs (x, x^+), (x, x^-) are formed from neutralization assays distinguishing "similar" vs. "escape" variants. Empirical distortion audits on such paths support a small $\varepsilon(L_\star)$ and justify using Euclidean displacements as proxies for antigenic geodesics.
- In VIDAR, f_θ is an ArcFace-style encoder whose outputs lie on the hypersphere $\mathbb{S}^{d-1}$; $\mathcal{P}(L)$ consists of short identity trajectories (arcs) with bounded geodesic length; and contrastive structure comes from same- vs. different-identity pairs. Small-angle regimes on $\mathbb{S}^{d-1}$ provide a natural bounded-distortion region, and temporal-smoothness regularization further reduces $K(\lambda)$.

In both cases, the existence theorem explains *why* the Euclidean computations used at runtime are meaningful: they operate inside a Davis manifold whose geometry has been constructed to reflect the relevant notion of "identity" or "configuration" within a bounded-distortion regime.

4 Detection Guarantees for Davis Systems

Section **??** showed how contrastive training with smoothness regularization can construct Davis manifolds with a bounded-distortion regime and nontrivial configuration margins. We now turn to *traversal*: how statistic-level separation on such a system can be translated into finite-variance risk bounds and an end-to-end correctness guarantee for high-risk alerts, together with an explicit error budget.

Throughout this section we fix a Davis system

$$\mathsf{D} = (\mathcal{X}, \mathcal{M}, g, \varphi, \psi, c, h, \mathcal{P}(L_\star), S)$$

satisfying Assumptions A1–A6 from Section **??** on an in-distribution region $\Omega_{\text{in}} \subset \mathcal{X}$, and we restrict attention to non-abstaining inputs (i.e., clips/samples for which all quality and OOD gates pass).

We write $Y \in \{0,1\}$ for the coarsened label, $Y = 1$ for the high-risk class (e.g., escape, synthetic, anomalous), with class priors $\pi_y = \mathbb{P}(Y = y)$, and let

$$S = S(X) \in \mathbb{R}$$

denote the scalar detection statistic produced by the system for an input X.

4.1 From Geometric Margins to Statistic Separation

The existence result (Theorem **??**) ensures that, on Ω_{in}, Davis manifolds admit (i) a bounded-distortion path regime with radius $\varepsilon_\star = \varepsilon(L_\star)$ and (ii) hard/soft configuration margins $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$ with an explicit non-vacuity gap. We now show how these geometric quantities induce separation at the level of the scalar statistic S under the monotonicity assumptions on features.

Recall that in a Davis system the detection statistic is constructed as

$$S = w^\top F + b, \qquad F_k = \phi_k(D), \quad k = 1, \dots, K,$$

where D is some scalar path-level or distance-like quantity on $\mathcal{M}$ (e.g., an aggregated geodesic or Euclidean distance along paths in $\mathcal{P}(L_\star)$), the feature maps $\phi_k : \mathbb{R}_+ \to \mathbb{R}$ are nondecreasing with derivative bounded below on an operational range, and the stacker weights $w_k \geq 0$.

Proposition 2 (Geometric margin $\Rightarrow$ statistic separation)**.** *Let* D *be a Davis system satisfying A1–A5. Assume:*

(a) (Soft configuration margin) For non-ambiguous configurations $h(c(z)) \in \{0,1\}$ *we have*

$$h(c(z)) = 0 \Rightarrow d_g(z, z^0) \leq \kappa_{\text{hard}}, \qquad h(c(z)) = 1 \Rightarrow d_g(z, z^1) \geq \kappa_{\text{soft}}$$

for some reference points or sets $z^0, z^1 \in \mathcal{M}$ *and* $\kappa_{\text{soft}} > \kappa_{\text{hard}} \geq 0$.

(b) (Bounded distortion on paths) For all paths $\gamma \in \mathcal{P}(L_\star)$ *and points* $z, z' \in \gamma([0,1])$,

$$(1 - \varepsilon_\star) d_g(z, z') \leq \delta\big(\psi(z), \psi(z')\big) \leq (1 + \varepsilon_\star) d_g(z, z').$$

(c) (Feature monotonicity) There exist constants $m_k > 0$ *and an operational range* $[d_{\min}, d_{\max}]$ *such that for all* $d \in [d_{\min}, d_{\max}]$,

$$\phi_k'(d) \geq m_k, \qquad k = 1, \dots, K,$$

and $D(X) \in [d_{\min}, d_{\max}]$ *for all non-abstaining* X.

(d) (Nonnegative fusion) The stacker weights satisfy $w_k \geq 0$.

Then there exists $\alpha > 0$*, depending only on* (w_k, m_k)*, such that the class-conditional means* $\mu_y := \mathbb{E}[S \mid Y = y]$ *satisfy*

$$\Delta_S := \mu_1 - \mu_0 \ \ \geq \ \ \alpha(1 - \varepsilon_\star)\left(\kappa_{\text{soft}} - \kappa_{\text{hard}}\right)_+,$$

where $(\cdot)_+ = \max\{\cdot, 0\}$. *In particular, whenever the non-vacuity condition* $\kappa_{\text{soft}} > \kappa_{\text{hard}}$ *holds and* $\varepsilon_\star < 1$, *we have* $\Delta_S > 0$.

Proof sketch. By the configuration margin (a), any two non-ambiguous points z_0, z_1 with different coarse labels satisfy $d_g(z_0, z_1) \geq \kappa_{\text{soft}} - \kappa_{\text{hard}}$ along some path in $\mathcal{P}(L_\star)$ (or along a concatenation of such paths). Bounded distortion (b) then implies that for the associated distance-like quantity D we have

$$D_1 - D_0 \;\gtrsim\; (1 - \varepsilon_\star)\left(\kappa_{\text{soft}} - \kappa_{\text{hard}}\right)_+,$$

up to constants that can be absorbed into the definition of α.

Feature monotonicity (c) and nonnegative fusion (d) imply that for any $D_1 > D_0$ in the operational range,

$$S(D_1) - S(D_0) = \sum_{k=1}^{K} w_k(\phi_k(D_1) - \phi_k(D_0)) \;\geq\; \Big(\sum_{k=1}^{K} w_k m_k\Big)(D_1 - D_0).$$

Taking expectations over $Y = 1$ and $Y = 0$ and combining with the geometric gap above yields the stated lower bound with $\alpha := \sum_k w_k m_k$. A full derivation, including explicitly tracking constants and conditioning on the ambiguous band $[\kappa_{\text{hard}}, \kappa_{\text{soft}}]$, appears in Appendix B. □

Intuition. The Davis construction ensures that different configurations are geodesically separated; bounded distortion says Euclidean distances see (almost) the same gap; feature monotonicity and nonnegative fusion ensure that larger distances monotonically increase the scalar statistic. Proposition **??** formalizes the chain

$$\text{configuration margin} \Rightarrow \text{geodesic gap} \Rightarrow \text{Euclidean gap} \Rightarrow \text{statistic gap}.$$

In practice, we will take Δ_S as an empirically estimated quantity, but Proposition **??** explains why geometry plus monotone features cannot yield $\Delta_S \approx 0$ unless either the margin collapses or the distortion bound is violated.

4.2 Finite-Variance Risk Bounds for a Scalar Statistic

Given a Davis system with statistic-level separation $\Delta_S > 0$, we next derive a finite-variance bound on misclassification probabilities using a Cantelli inequality. This mirrors the HERALD and VIDAR analyses, but we keep notation abstract.

Let

$$\mu_y = \mathbb{E}[S \mid Y = y], \qquad \sigma_y^2 = \mathrm{Var}(S \mid Y = y), \qquad y \in \{0, 1\},$$

and $\Delta_S = \mu_1 - \mu_0 > 0$. Consider the midpoint threshold

$$s_\star := \frac{\mu_0 + \mu_1}{2} = \mu_0 + \frac{\Delta_S}{2}.$$

Proposition 3 (Cantelli bound for Davis statistics)**.** *Assume A1 (finite variance) and* $\Delta_S > 0$. *Define one-sided Cantelli tails*

$$c_0 := \frac{\sigma_0^2}{\sigma_0^2 + (\Delta_S/2)^2}, \qquad c_1 := \frac{\sigma_1^2}{\sigma_1^2 + (\Delta_S/2)^2},$$

and let $\pi_y = \mathbb{P}(Y = y)$. Then:

$$\mathbb{P}(S > s_\star \mid Y = 0) \le c_0,$$
$$\mathbb{P}(S \le s_\star \mid Y = 1) \le c_1,$$

and the posterior correctness of high-score points satisfies

$$\mathbb{P}(Y = 1 \mid S > s_\star) \;\ge\; \frac{\pi_1(1 - c_1)}{\pi_0 c_0 + \pi_1(1 - c_1)}.$$

When $(\Delta_S/2)^2 \gg \sigma_0^2, \sigma_1^2$, both c_0 and c_1 are small and the lower bound is close to one.

Proof. Cantelli's inequality states that for any real-valued random variable X with mean μ and variance $\sigma^2 < \infty$ and any $a > 0$,

$$\mathbb{P}(X - \mu \ge a) \le \frac{\sigma^2}{\sigma^2 + a^2}.$$

Apply this with $X = S \mid Y = 0$, $\mu = \mu_0$ and $a = \Delta_S/2$ to obtain the bound on $\mathbb{P}(S > s_\star \mid Y = 0)$, and with $X = -S \mid Y = 1$ (so that $X - \mu = \mu_1 - S$) to bound $\mathbb{P}(S \le s_\star \mid Y = 1)$.

For the posterior, Bayes' rule gives

$$\mathbb{P}(Y = 1 \mid S > s_\star) = \frac{\pi_1 \mathbb{P}(S > s_\star \mid Y = 1)}{\pi_0 \mathbb{P}(S > s_\star \mid Y = 0) + \pi_1 \mathbb{P}(S > s_\star \mid Y = 1)}.$$

Using $\mathbb{P}(S > s_\star \mid Y = 0) \le c_0$ and $\mathbb{P}(S > s_\star \mid Y = 1) \ge 1 - c_1$ yields the stated fraction. □

Remark (Finite-variance vs. sub-Gaussian). Proposition **??** requires only finite second moments (A1); it does not assume sub-Gaussian tails. Sub-Gaussian assumptions would yield exponentially decaying bounds in Δ_S^2, but are often unrealistic for heavy-tailed statistics. Cantelli instead offers a robust, polynomial-decay bound that matches the HERALD and VIDAR finite-variance guarantees.

For later use, we define the *Cantelli term*

$$B_{\text{Cantelli}}(\Delta_S, \sigma_0, \sigma_1, \pi_0, \pi_1) := \frac{\pi_1(1 - c_1)}{\pi_0 c_0 + \pi_1(1 - c_1)}.$$

In practice, $(\mu_y, \sigma_y^2, \pi_y)$ are estimated on a held-out validation set, and an additional estimation slack $\varepsilon_{\text{stat}}$ is subtracted from the right-hand side; we return to this in Theorem **??**.

4.3 Error Decomposition for Davis Systems

The Cantelli bound quantifies how separation in S controls misclassification *conditional* on using the correct statistic. To obtain a system-level guarantee for high-risk alerts, we must account for several distinct sources of error.

Let H denote the event that the system issues a `high_risk` decision on an input X, under some operational threshold τ for S and auxiliary gates (Section **??** in the outline). Let C denote the event that this decision is correct, i.e. $Y = 1$.

We decompose error into four components:

- **Geometry error E_{geom}.** Probability that the Davis manifold geometry behaves outside its validated regime on a non-abstaining input—e.g., bounded-distortion or smoothness assumptions fail, or the path-class choice $\mathcal{P}(L_\star)$ does not capture the actual transformation, in a way that materially corrupts S.
- **Linkage error E_{link}.** Probability that the detection pipeline linking geometry to S fails even when geometry is good—e.g., feature extraction breaks, the PIT nulls F_k are misspecified, or the learned fusion fails to preserve the monotone relationship between distance and event indicator.
- **Calibration error ξ.** Discrepancy between the calibrated output and the true conditional probability $\mathbb{P}(Y = 1 \mid S)$ in the high-risk region, e.g. measured by an expected calibration error (ECE) or Brier decomposition restricted to high scores.
- **Abstention failure ζ.** Probability that the system *does not* abstain when it should—e.g., when distortion audits, feature-range checks, or support conditions indicate that the Davis assumptions are violated, but a non-abstaining prediction is nonetheless issued.

Let G_{geom}, G_{link}, G_{cal}, and G_{abst} denote the complements of these error events (geometry behaves, linkage is faithful, calibration is adequate, abstention policy behaves), and write

$$G := G_{\text{geom}} \cap G_{\text{link}} \cap G_{\text{cal}} \cap G_{\text{abst}}.$$

By definition,

$$\mathbb{P}(G^c_{\text{geom}}) = E_{\text{geom}}, \quad \mathbb{P}(G^c_{\text{link}}) = E_{\text{link}}, \quad \mathbb{P}(G^c_{\text{cal}}) = \xi, \quad \mathbb{P}(G^c_{\text{abst}}) = \zeta.$$

We will also use the shorthand

$$E_{\text{sum}} := E_{\text{geom}} + E_{\text{link}} + \xi + \zeta.$$

Assumption A6 asserts that these errors are approximately independent after conditioning on being in-regime and non-abstaining, or at least that their joint failure probability is close to the product of their marginals. To capture deviations from independence, we introduce an *independence slack* term:

Definition 10 (Independence slack). *The independence slack $\delta_{\text{indep}} \in [0, 1]$ is defined as*

$$\delta_{\text{indep}} \;:=\; \mathbb{P}(G^c) - \big(1 - (1 - E_{\text{geom}})(1 - E_{\text{link}})(1 - \xi)(1 - \zeta)\big)_+.$$

Equivalently, δ_{indep} measures how much more often some *component fails than would be expected under independence.*

When the four error sources are genuinely independent on the evaluation distribution, $\delta_{\text{indep}} \approx 0$; when they correlate (e.g., geometry and linkage both fail when curvature is high), δ_{indep} is positive and must be estimated empirically.

4.4 Main Compositional Guarantee

We are now ready to state the main theorem: a lower bound on the correctness of high-risk alerts that combines the Cantelli term with the Davis error budget.

Let τ denote the operational high-risk threshold on S, and assume it is chosen so that

$$H \subseteq \{ X : S(X) > s_\star \};$$

in practice, it is common to take $\tau \geq s_\star$ plus additional margin.

Theorem 2 (Main Davis bound)**.** *Let* D *be a Davis system satisfying Assumptions A1–A6 on an in-distribution region* Ω_{in}*, with path horizon* $L_\star$*, distortion bound* $\varepsilon_\star$*, and configuration margins* $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$ *satisfying the non-vacuity condition*

$$\kappa_{\text{soft}} - \kappa_{\text{hard}} - \varepsilon_\star L_\star > 0.$$

Assume:

(i) $\Delta_S = \mu_1 - \mu_0 > 0$ *on in-regime, non-abstaining inputs, with* $(\mu_y, \sigma_y^2, \pi_y)$ *finite and estimated from validation data;*

(ii) the high-risk event H *implies* $S > s_\star$ *and no abstention;*

(iii) the system exhibits positive correlation between alerts and good behavior: $\mathbb{P}(G \mid H) \geq \mathbb{P}(G)$*, meaning that high-risk alerts are at least as likely to originate from in-regime operation as from degraded states;*

(iv) the error components $(E_{\text{geom}}, E_{\text{link}}, \xi, \zeta)$ *and independence slack* δ_{indep} *are estimated on held-out data, with aggregate statistical estimation error* $\varepsilon_{\text{est}} \geq 0$*.*

Then the posterior correctness of high-risk alerts satisfies

$$\mathbb{P}(Y = 1 \mid H) \;\geq\; \Big[(1 - E_{\text{geom}})(1 - E_{\text{link}})(1 - \xi)(1 - \zeta) - \delta_{\text{indep}}\Big]_+ B_{\text{Cantelli}}(\Delta_S, \sigma_0, \sigma_1, \pi_0, \pi_1) \;-\; \varepsilon_{\text{est}}, \tag{6}$$

where $[u]_+ := \max\{u, 0\}$ *and* B_{Cantelli} *is as in Proposition* **??***. In particular, when all error terms and the Cantelli tails* c_0, c_1 *are small, the right-hand side is meaningfully above zero; when they are large, the bound becomes vacuous and the Davis system should not be relied on in that regime.*

Proof sketch. Condition on the "good" event G (geometry, linkage, calibration, and abstention all behave). On G and H, the Davis assumptions ensure:

- the geometry is within the bounded-distortion regime and configuration margins apply (A2, A5);
- the feature construction preserves distance separation and the learned fusion behaves monotonically (A3, Def. **??**);
- the scalar statistic S has separation $\Delta_S > 0$ and finite variance (A1);
- calibration is well-behaved in the high-risk region and abstention has removed obvious out-of-regime points.

Thus Proposition **??** applies to the conditional distribution of (S, Y) under G, giving

$$\mathbb{P}(Y = 1 \mid S > s_\star, G) \;\geq\; B_{\text{Cantelli}}(\Delta_S, \sigma_0, \sigma_1, \pi_0, \pi_1).$$

Next, relate H to $S > s_\star$: by assumption $H \subseteq \{S > s_\star\}$ and abstention is false on H, so

$$\mathbb{P}(Y = 1 \mid H, G) \;\geq\; \mathbb{P}(Y = 1 \mid S > s_\star, G) \;\geq\; B_{\text{Cantelli}}(\cdot).$$

Finally, decompose over G:

$$\mathbb{P}(Y = 1 \mid H) = \mathbb{P}(Y = 1 \mid H, G)\,\mathbb{P}(G \mid H) + \mathbb{P}(Y = 1 \mid H, G^c)\,\mathbb{P}(G^c \mid H).$$

Since $\mathbb{P}(Y = 1 \mid H, G^c) \geq 0$,

$$\mathbb{P}(Y = 1 \mid H) \;\geq\; B_{\text{Cantelli}}(\cdot)\,\mathbb{P}(G \mid H).$$

Lower-bound $\mathbb{P}(G \mid H)$ by $1 - \mathbb{P}(G^c)$ and use

$$\mathbb{P}(G^c) \leq E_{\text{geom}} + E_{\text{link}} + \xi + \zeta \quad \text{(union bound)},$$

or, under approximate independence, by the multiplicative form minus δ_{indep},

$$\mathbb{P}(G) \geq (1 - E_{\text{geom}})(1 - E_{\text{link}})(1 - \xi)(1 - \zeta) - \delta_{\text{indep}}.$$

The $[\cdot]_+$ truncation prevents negative lower bounds in regimes where E_{sum} is large. Replacing population quantities by their empirical estimates and applying standard concentration inequalities yields the additional slack ε_{est}. Full details are given in Appendix C. □

Interpretation. The bound (**??**) decomposes into three factors:

- *Stability–sensitivity factor*

$$(1 - E_{\text{geom}})(1 - E_{\text{link}})(1 - \xi)(1 - \zeta) - \delta_{\text{indep}}$$

  quantifies how often the Davis system is "well-behaved": geometry within the validated regime, features and fusion correctly reflecting geometry, calibration accurate in the high-risk region, and abstention correctly triggered when assumptions fail. The slack δ_{indep} clips away optimistic gains from assuming independence when errors are correlated.

- *Cantelli term*

$$B_{\text{Cantelli}}(\Delta_S, \sigma_0, \sigma_1, \pi_0, \pi_1) = \frac{\pi_1(1 - c_1)}{\pi_0 c_0 + \pi_1(1 - c_1)}$$

  expresses how much class-conditional separation in the statistic S (through Δ_S and the variances) translates into posterior confidence on high scores, under finite variance.

- *Estimation slack* ε_{est} accounts for finite-sample uncertainty in all estimated quantities.

In words: high-risk alerts are correct whenever (i) the Davis construction behaves as validated and (ii) S meaningfully separates classes; Theorem **??** makes this precise and quantifies how violations in either channel degrade the guarantee.

4.5 Vacuity, Fallbacks, and Domain Instantiations

The bound in Theorem **??** is deliberately conservative. It also makes clear when it becomes *vacuous* and the system should abstain or be interpreted only heuristically.

Corollary 1 (Vacuity threshold)**.** *Suppose $E_{\text{sum}} = E_{\text{geom}} + E_{\text{link}} + \xi + \zeta$ is estimated on a held-out validation set. If*

$$E_{\text{sum}} > \tau_{\text{vac}} \qquad \textit{for some pre-registered } \tau_{\text{vac}} \in (0,1),$$

then

$$\mathbb{P}(Y = 1 \mid H) \approx 0 \quad \textit{(lower bound)}$$

for any reasonable choice of δ_{indep}, and the Davis system should treat the bound as non-informative in that regime. In particular, when $E_{\text{sum}} > 1$ the union-bound fallback is identically zero or negative and must be truncated.

A practical choice is $\tau_{\text{vac}} \approx 0.7$: when the sum of error-budget components exceeds 0.7, the multiplicative and union-bound forms both yield weak guarantees, and abstention or deferral is preferable to exuding unwarranted confidence.

HERALD and VIDAR as Davis systems. In subsequent sections we instantiate Theorem **??** for:

- HERALD, where S is an antigenic drift statistic $D(t)$, E_{geom} captures violations of the pullback antigenic manifold, E_{link} covers misalignment between drift and escape, and ξ, ζ correspond to calibration and abstention in epidemiological surveillance.
- VIDAR, where S is a deepfake detection statistic built from Riemannian identity manifold features and auxiliary detectors, E_{geom} covers failures of the identity geometry on S^{d-1}, E_{link} covers feature linkage failures, and ξ, ζ again represent calibration and abstention error for forensic decisions.

In both cases, the abstract parameters $(\Delta_S, \sigma_0, \sigma_1, \pi_y, E., \delta_{\text{indep}})$ map to concrete, empirically estimable quantities described in the corresponding domain sections and appendices.

Section **??** will complement Theorem **??** by studying how the path horizon $L_\star$ and distortion profile $\varepsilon(L)$ interact with separation Δ_S, and how to choose $L_\star$ to balance coverage against geometric stability.

5 Path-Class Design and Optimal Path Horizon

The guarantees in Sections **??** and **??** are explicitly conditional on a *path family* $\mathcal{P}(L)$ and a chosen path horizon $L_\star$. In practice, the choice of $\mathcal{P}(L)$ and $L_\star$ is as important as the choice of encoder or regularization strength: it controls which semantic changes are detectable, which regimes of bounded distortion are audited, and whether the main bound of Theorem 2 is informative or vacuous.

This section formalizes the role of path families (Section **??**), describes how the Davis bound acquires an explicit L-dependence (Section **??**), and gives a toy optimization example illustrating the existence of a non-trivial optimal horizon $L_\star$ (Section **??**). We then distill practical guidelines for choosing $\mathcal{P}(L_\star)$ in HERALD- and VIDAR-like systems (Section **??**) and summarize limitations (Section **??**).

5.1 The role of the path family

Recall from Definition **??** that a (piecewise) smooth path is a map $\gamma : [0,1] \to \mathcal{M}$ with geodesic length

$$\ell_g(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|_g \, dt,$$

and that a path family $\mathcal{P}(L)$ consists of all such paths of length at most L that satisfy a domain-specific admissibility criterion (e.g., epitope-restricted mutation paths in HERALD, or smooth identity trajectories on S^{d-1} in VIDAR).

At a high level, $\mathcal{P}(L)$ plays three roles:

(R1) **Regime of validity.** The bounded-distortion condition in Proposition **??** is stated for paths whose length and curvature are controlled. For each regularization strength λ, there exist constants $K(\lambda)$ and $C' > 0$ such that, for all $\gamma \in \mathcal{P}(L)$ lying in an in-distribution region of radius R,

$$\varepsilon(L) \;\le\; C' K(\lambda) L, \tag{7}$$

where $\varepsilon(L)$ is the smallest distortion radius such that geodesic–Euclidean distortion along γ is bounded by $\varepsilon(L)$ in the sense of (**??**). Longer paths and weaker smoothness (smaller λ) both increase $\varepsilon(L)$.

(R2) **Detection scope.** The probability that a semantically meaningful change is *reachable and visible* within horizon L,

$$P_{\text{detect}}(L) \;=\; \mathbb{P}\big[\exists\, \gamma \in \mathcal{P}(L) \text{ that leaves the current configuration region within horizon } L\big],$$

is a non-decreasing function of L. Enlarging $\mathcal{P}(L)$—by allowing more mutation sites, longer arcs, or additional transformations—increases the fraction of real events that can manifest as configuration changes along some $\gamma \in \mathcal{P}(L)$.

(R3) **Margin survival.** The configuration margins in Definition **??** are preserved only if distortion does not "eat away" the soft margin. A non-vacuity condition of the form

$$\kappa_{\text{soft}} - 2R\varepsilon(L_\star) \;>\; 0 \tag{8}$$

guarantees that points in the same configuration ball of radius R remain closer to each other than to points across the configuration boundary, after accounting for geodesic–Euclidean distortion along all paths in $\mathcal{P}(L_\star)$.

Thus, $\mathcal{P}(L)$ and $L_\star$ jointly control a three-way trade-off:

$$\text{coverage} \quad \text{vs.} \quad \text{geometric fidelity} \quad \text{vs.} \quad \text{margin survival.}$$

Sections **??** and **??** make this trade-off more explicit.

5.2 How the Davis bound depends on L

Theorem 2 in Section **??** provides a lower bound

$$P(Y = 1 \mid H) \;\ge\; \text{LB}_{\text{Davis}},$$

for high-risk events H (alerts or dominance events, depending on instantiation), where $\mathrm{LB}_{\mathrm{Davis}}$ factors into:

- a *stability–sensitivity factor* depending on $(E_{\mathrm{geom}}, E_{\mathrm{link}}, \xi, \zeta, \delta_{\mathrm{indep}})$, and
- a *Cantelli term* depending on the separation Δ_S and variances of the detection statistic.

Both pieces depend implicitly on L through the distortion bound $\varepsilon(L)$ and the path family $\mathcal{P}(L)$. For intuition, we make this dependence explicit.

From distortion to statistic separation. Under the bounded-distortion regime of Proposition **??**, for any pair of points (z, z') that are endpoints of a path in $\mathcal{P}(L)$, we have

$$(1-\varepsilon(L))d_g(z,z') \;\le\; \delta(z,z') \;\le\; (1+\varepsilon(L))d_g(z,z').$$

If the features $\{F_k\}$ used to construct the statistic S are monotone in δ with slopes bounded below by $m_k > 0$ on the operational range, and if the fusion layer is linear with weights $\{w_k\}$, then the separation result in Section **??** yields

$$\Delta_S(L) \;\ge\; \alpha(1-\varepsilon(L))\Delta_g(L), \qquad \alpha \;=\; \sum_k w_k m_k, \tag{9}$$

where $\Delta_g(L)$ is the geodesic separation between configurations induced along paths in $\mathcal{P}(L)$.

Two channels of L-dependence appear here:

(a) $\varepsilon(L)$ typically increases with L by (**??**), shrinking the effective separation;

(b) $\Delta_g(L)$ may increase with L because longer paths allow more substantial configuration changes.

For fixed construction (λ and $K(\lambda)$), $\varepsilon(L)$ captures how far along a path one can traverse while remaining in the well-approximated, low-curvature regime.

From separation to the Cantelli term. The Cantelli component of Theorem 2 is a function of $\Delta_S(L)$ and the class-conditional variances $\sigma_y^2(L)$:

$$\mathrm{Cantelli}(L) \;=\; \frac{\pi_1(1-c_1(L))}{\pi_0 c_0(L) + \pi_1(1-c_1(L))}, \quad c_y(L) \;=\; \frac{\sigma_y^2(L)}{\sigma_y^2(L) + (\Delta_S(L)/2)^2}.$$

As $\Delta_S(L)$ increases relative to $\sigma_y(L)$, both $c_0(L)$ and $c_1(L)$ decrease and the Cantelli term improves; as $\Delta_S(L)$ decreases, the term collapses toward the class prior baseline.

Combining (**??**) with the Cantelli expression shows that, to first order, the effect of L on this component is governed by the product

$$\bigl(1-\varepsilon(L)\bigr)^2 \bigl(\Delta_g(L)\bigr)^2.$$

Detection power and coverage. The second way L enters the Davis bound is through the probability that relevant events are *detectable* under $\mathcal{P}(L)$. For a given application, we can define an *operational detection power* at horizon L,

$$P_{\mathrm{detect}}(L) \;=\; \mathbb{P}\bigl[\exists \gamma \in \mathcal{P}(L) \text{ along which the configuration changes and the system does not abstain}\bigr].$$

Increasing L enlarges $\mathcal{P}(L)$ and therefore $P_{\text{detect}}(L)$, but also enlarges $\varepsilon(L)$ and pushes more paths toward the edge of the validated regime where geometry and features may become unreliable, increasing E_{geom} and ζ.

Putting it together. Abstractly, we can think of the Davis lower bound as a function of L of the form

$$\mathrm{LB}_{\text{Davis}}(L) \;\approx\; P_{\text{detect}}(L) \times \underbrace{\left[(1-E_{\text{geom}}(L))(1-E_{\text{link}}(L))(1-\xi)(1-\zeta(L)) - \delta_{\text{indep}}\right]}_{\text{stability–sensitivity factor}} \times \mathrm{Cantelli}(\Delta_S(L), \sigma_0(L), \sigma_1(L)) \tag{10}$$

with $\Delta_S(L)$ constrained by (**??**) and $\varepsilon(L)$ constrained by (**??**). Section **??** explores this dependence in a simplified one-dimensional model to show that non-trivial optima $L_\star$ naturally arise.

5.3 A toy model for the optimal path horizon

To make the trade-offs in (**??**) concrete, we consider a caricature model in which:

- distortion grows linearly with path length: $\varepsilon(L) = aL$ for some $a > 0$;
- geodesic separation Δ_g is approximately constant over the range of interest (i.e., the dominant effect of L is distortion rather than changing which configurations are reachable);
- detection power saturates with L according to

  $$P_{\text{detect}}(L) = 1 - e^{-L/L_0},$$

  with a characteristic length scale $L_0 > 0$; and
- the Cantelli term is driven by the squared effective separation $\Delta_S(L)^2 \propto (1-\varepsilon(L))^2$, with variances and error-budget terms held fixed.

Under these simplifications, a normalized surrogate for the Davis bound is

$$\widetilde{\mathrm{LB}}(L) \;=\; P_{\text{detect}}(L) \times \frac{\Delta_S(L)^2}{\Delta_S(L)^2 + \sigma^2} \;\propto\; \left(1 - e^{-L/L_0}\right) \cdot \frac{(1-aL)^2}{(1-aL)^2 + \sigma^2}, \tag{11}$$

for L such that $aL < 1$ (i.e., before the distortion bound becomes useless).

Example 5.1 (Illustrative optimization of $L_\star$). Fix units so that L is measured in "elementary" steps (amino-acid substitutions in HERALD, radians of arc length on S^{d-1} in VIDAR). Consider the following illustrative parameters:

$$a = 0.05, \qquad L_0 = 6, \qquad \sigma^2 = 1.$$

Then

$$\varepsilon(L) = 0.05L, \qquad P_{\text{detect}}(L) = 1 - e^{-L/6}.$$

For several representative horizons:

- $L = 3$: $\varepsilon(3) = 0.15$, so $1 - \varepsilon(3) = 0.85$. The detection power is $P_{\text{detect}}(3) \approx 1 - e^{-0.5} \approx 0.39$.

The squared separation factor is $\frac{0.85^2}{0.85^2+1} \approx \frac{0.72}{1.72} \approx 0.42$. Thus $\widetilde{\mathrm{LB}}(3) \propto 0.39 \times 0.42 \approx 0.16$ (arbitrary units).

- $L = 6$: $\varepsilon(6) = 0.30$, so $1 - \varepsilon(6) = 0.70$. Detection power is $P_{\text{detect}}(6) \approx 1 - e^{-1} \approx 0.63$. The squared separation factor is $\frac{0.70^2}{0.70^2+1} \approx \frac{0.49}{1.49} \approx 0.33$. Thus $\widetilde{\mathrm{LB}}(6) \propto 0.63 \times 0.33 \approx 0.21$.
- $L = 10$: $\varepsilon(10) = 0.50$, so $1 - \varepsilon(10) = 0.50$. Detection power is $P_{\text{detect}}(10) \approx 1 - e^{-10/6} \approx 0.81$. The squared separation factor is $\frac{0.50^2}{0.50^2+1} = \frac{0.25}{1.25} = 0.20$. Thus $\widetilde{\mathrm{LB}}(10) \propto 0.81 \times 0.20 \approx 0.16$.
- $L = 15$: $\varepsilon(15) = 0.75$, so $1 - \varepsilon(15) = 0.25$. Detection power is $P_{\text{detect}}(15) \approx 1 - e^{-15/6} \approx 0.92$. The squared separation factor is $\frac{0.25^2}{0.25^2+1} = \frac{0.0625}{1.0625} \approx 0.059$. Thus $\widetilde{\mathrm{LB}}(15) \propto 0.92 \times 0.059 \approx 0.054$.

In this cartoon, $\widetilde{\mathrm{LB}}(L)$ increases from $L = 3$ to a peak near $L \approx 6$ and then degrades as distortion overwhelms the benefit of additional detection power. The precise numerical values are not important; the qualitative shape (a single interior maximum) is robust across reasonable choices of (a, L_0, σ^2) with aL_0 neither too small nor too large. Figure **??** plots a typical instance of (**??**).

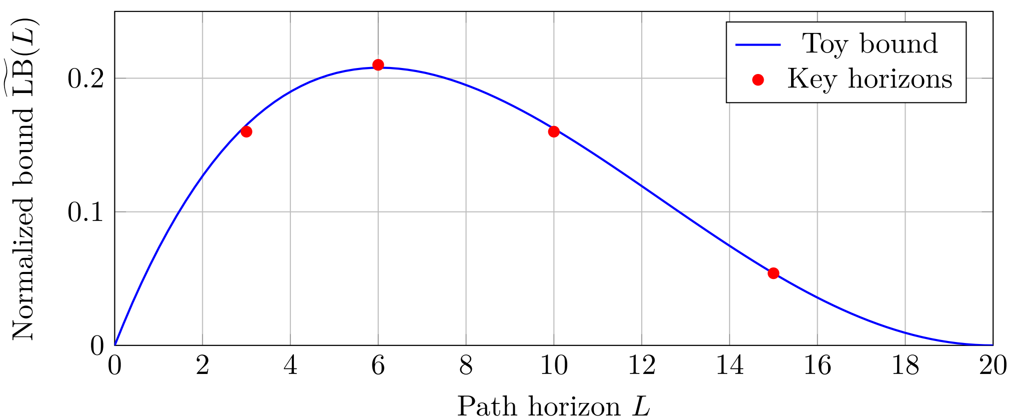

Figure 2: **Trade-off between path horizon and detection bound (toy model).** The normalized lower bound $\widetilde{\mathrm{LB}}(L)$ from Equation (**??**) exhibits a single interior maximum at $L_\star \approx 6$, balancing detection power $P_{\text{detect}}(L)$ (which increases with L) against distortion $\varepsilon(L)$ (which also increases with L). Parameters: $a = 0.05$, $L_0 = 6$, $\sigma^2 = 1$.

This example illustrates why the path horizon $L_\star$ should be treated as a tunable parameter rather than fixed *a priori*: too small and the system misses many events (low P_{detect}); too large and the geometric approximation degrades (large $\varepsilon(L)$), shrinking the effective separation and pushing the Davis bound toward vacuity.

5.4 Guidelines for choosing $\mathcal{P}(L_\star)$ in practice

We now translate the abstract trade-offs above into concrete guidelines for two representative domains, using the parameter ranges in Section **??** as reference.

HERALD-style antigenic drift. In HERALD-like settings, a natural path family is the class of mutation paths restricted to epitope positions:

$$\mathcal{P}_{\text{HER}}(L) \;=\; \Big\{\gamma \text{ induced by at most } L \text{ single-amino-acid substitutions in a pre-specified epitope set}\Big\}.$$

Guidelines:

- *Start with the biological horizon.* Survey historical antigenic drift to identify the typical number of substitutions associated with meaningful immune escape (e.g., 3–10 changes in key epitopes over a season). This suggests a plausible range for $L_\star$.

- *Audit distortion vs. L.* For candidate horizons (e.g., $L \in \{3, 5, 8, 10\}$), simulate or replay mutational paths in $\mathcal{P}_{\text{HER}}(L)$ and measure empirical distortion ratios along each path. Choose $L_\star$ such that

$$\varepsilon(L_\star) \le \varepsilon_{\text{target}},$$

  where $\varepsilon_{\text{target}}$ is set by a non-vacuity condition of the form $\kappa_{\text{soft}} - 2R\varepsilon(L_\star) > 0$ and by the domain's tolerance for geometric error.

- *Estimate detection power.* Using retrospective data or simulations, estimate $P_{\text{detect}}(L)$ as the fraction of historically important trajectories whose relevant configuration changes can be realized within $\mathcal{P}_{\text{HER}}(L)$. This empirically populates the toy trade-off in (**??**).

- *Select $L_\star$ via validation.* For each candidate L, instantiate the Davis system, estimate the components of (**??**) on held-out data (separation $\Delta_S(L)$, error terms, coverage), and pick $L_\star$ that maximizes a surrogate of $\text{LB}_{\text{Davis}}(L)$ subject to $\varepsilon(L_\star)$ and error-budget constraints.

Empirically, horizons in the range $L_\star \approx 5$–10 mutations are plausible for seasonal respiratory viruses, but the framework does not assume any specific value.

VIDAR-style identity dynamics. In VIDAR-like settings on S^{d-1}, the path family consists of short arcs traced by identity trajectories:

$$\mathcal{P}_{\text{VID}}(L) \;=\; \Big\{\gamma \text{ with arc length } \ell_g(\gamma) \le L \text{ along an identity trajectory on } S^{d-1}\Big\}.$$

Guidelines:

- *Anchor L in typical motion.* Measure frame-to-frame geodesic distances along real identity trajectories to estimate typical per-frame angles and the distribution of cumulative arc lengths over windows. This yields a natural scale for L (e.g., 0.1–0.5 radians for short clips).

- *Check the small-angle regime.* Distortion and the validity of tangent-space approximations depend on remaining in a small-angle regime. Distortion audits analogous to Section **??** should verify that $\varepsilon(L)$ remains below a target threshold on real trajectories up to $L_\star$.

- *Tie to feature behavior.* Evaluate how RIM features (F1–F4) behave as functions of L; for example, whether subspace leakage (F4) starts to saturate beyond a certain arc length, indicating diminishing returns from larger L.

- *Tune $L_\star$ on validation clips.* As in the HERALD-style case, instantiate the Davis system for several candidate L values, estimate the Davis bound components, and pick $L_\star$ that balances detection power (ability to see identity anomalies) with geometric fidelity and error-budget constraints.

In both domains, the key point is that $L_\star$ is not a free, purely heuristic parameter: it is constrained by distortion audits, margin non-vacuity, and empirical detection power.

5.5 Limitations of the path-optimization view

The toy analysis above is intentionally simplified. In real systems:

- The distortion function $\varepsilon(L)$ may deviate from linearity (e.g., saturating for very short paths or accelerating beyond a curvature threshold).
- The geodesic separation $\Delta_g(L)$ is not strictly independent of L; enlarging $\mathcal{P}(L)$ can change which configurations are reachable and how often they occur.
- Detection power $P_{\text{detect}}(L)$ is shaped by domain-specific constraints (e.g., epitope masks, occlusions, motion patterns) that may not follow a simple exponential saturation.
- Error terms $E_{\text{geom}}(L)$ and $\zeta(L)$ may increase sharply once L crosses a regime boundary, making the bound in Theorem 2 vacuous even if separation remains non-zero.
- Slices of the data (by demographic, capture condition, or generator family) may exhibit different optimal horizons $L_\star$, suggesting that a single global $L_\star$ is suboptimal.

For these reasons, Equation (**??**) should be viewed as a conceptual guide rather than a prescriptive formula. The practical recommendation is to: (i) parameterize $\mathcal{P}(L)$ in a domain-appropriate way; (ii) perform distortion audits and feature-behavior checks as L varies; (iii) estimate the components of the Davis bound on held-out data; and (iv) treat $L_\star$ as a hyperparameter chosen to maximize a *validated* lower bound, not merely an empirical score.

In summary, path-class design is where geometric theory, domain knowledge, and operational constraints meet. The Davis framework provides the language and structure for this choice; the actual selection of $\mathcal{P}(L_\star)$ is an empirical design decision that must be documented, audited, and revisited as data and deployment conditions evolve.

6 Operational Protocols and Validation

Sections **??**–**??** provide the theoretical foundation for Davis systems: existence guarantees from contrastive training and smoothness (Theorem **??**), finite-variance detection bounds (Theorem **??**), and path-horizon optimization. This section translates those results into concrete protocols for building, validating, and monitoring Davis systems in deployment.

The central operational challenge is that all Davis guarantees are *conditional*: they hold only when geometry, features, calibration, and abstention behave as assumed on an in-distribution region Ω_{in}. In practice, we cannot verify these assumptions exhaustively at training time, nor can we guarantee they persist in deployment. Instead, we must:

(i) **Audit** the bounded-distortion regime and feature behavior on held-out validation data;

(ii) **Estimate** each component of the error budget ($E_{\text{geom}}, E_{\text{link}}, \xi, \zeta, \delta_{\text{indep}}$) with confidence intervals;

(iii) **Monitor** these quantities during deployment and trigger abstention or retraining when they drift; and

(iv) **Document** all assumptions, thresholds, and procedures in a version-controlled dossier (Section **??**).

This section provides step-by-step protocols for each of these tasks, organized around the error-budget decomposition from Theorem **??**. Algorithm boxes summarize key procedures; tables provide decision rules; and worked examples show how HERALD and VIDAR instantiate these protocols.

6.1 Overview: The Validation Workflow

Table **??** (conceptual) summarizes the end-to-end validation workflow for a Davis system. The process begins with a trained encoder f_θ and proceeds through five phases:

Table 7: **Davis validation workflow (overview).** Each phase produces estimates or gates that feed into the error budget and abstention policies.

Phase	**Goal**	**Output**
1. Distortion audit	Validate $\varepsilon(L_\star) < \varepsilon_{\text{target}}$ on held-out paths	Empirical $\hat{\varepsilon}(L)$ curve, distortion gate thresholds
2. Margin estimation	Verify $\kappa_{\text{soft}} - 2R\hat{\varepsilon}(L_\star) > 0$	Estimates $(\hat{\kappa}_{\text{hard}}, \hat{\kappa}_{\text{soft}})$ with CIs
3. Feature validation	Check monotonicity and linkage on operational range	Feature slopes $\{m_k\}$, linkage failure rate $\hat{E}_{\text{link}}$
4. Calibration	Fit calibrator and measure ECE/Brier in high-risk region	Calibration map, $\hat{\xi}$ estimate
5. Error budget	Estimate all terms and independence slack	$(\hat{E}_{\text{geom}}, \hat{E}_{\text{link}}, \hat{\xi}, \hat{\zeta}, \hat{\delta}_{\text{indep}})$ and vacuity check

Data splits. All validation protocols assume access to three disjoint held-out sets:

- $\mathcal{D}_{\text{val}}$: primary validation set for distortion audits, margin estimation, and feature checks ($\approx 20\%$ of data);
- $\mathcal{D}_{\text{cal}}$: calibration set for fitting the calibrator and estimating ξ ($\approx 10\%$);
- $\mathcal{D}_{\text{test}}$: final error-budget estimation and integration test ($\approx 10\%$).

These should be stratified by any known slices (demographics, capture conditions, generator families) and temporally separated from training data when time-series structure is present.

6.2 Phase 1: Distortion Audits

The bounded-distortion assumption—that Euclidean distances approximate geodesic distances along paths in $\mathcal{P}(L_\star)$ within a factor $(1 \pm \varepsilon_\star)$—is the geometric foundation of Davis systems. Distortion audits empirically validate this assumption and identify regimes where it fails.

6.2.1 Protocol: Measuring Distortion Ratios

Algorithm 1 Distortion Audit on Validation Paths

Require: Encoder f_θ, path family $\mathcal{P}(L)$, validation set $\mathcal{D}_{\text{val}}$
Ensure: Empirical distortion profile $\hat{\varepsilon}(L)$, per-path distortion ratios $\{\rho_j\}$

```
Sample N_path paths {γ_j}_{j=1}^{N_path} from P(L_⋆) using D_val
for each path γ_j with waypoints (x_{j,1}, ..., x_{j,T_j}) do
    Embed: z_{j,t} = f_θ(x_{j,t}) for t = 1, ..., T_j
    Compute geodesic length: ℓ_g(γ_j) = Σ_{t=1}^{T_j−1} d_g(z_{j,t}, z_{j,t+1})        ▷ Using metric g
    Compute Euclidean length: ℓ_δ(γ_j) = Σ_{t=1}^{T_j−1} ‖z_{j,t} − z_{j,t+1}‖_2
    Distortion ratio: ρ_j = ℓ_δ(γ_j)/ℓ_g(γ_j)
end for
Compute summary statistics: ρ̄ = mean(ρ_j), σ_ρ = std(ρ_j), percentiles
Estimate distortion radius: ε̂(L_⋆) = max{|ρ_j − 1|} or 95th percentile of |ρ_j − 1|
return ε̂(L_⋆), {ρ_j}, slice-wise breakdowns
```

Implementation notes.

- **Geodesic approximation.** For general pullback manifolds, $d_g(z, z')$ is not available in closed form. Use:
  - *Riemannian metric integration*: Numerically integrate $\int \sqrt{g(\dot{\gamma}, \dot{\gamma})}\, dt$ along the path;
  - *Geodesic solver*: Run a Riemannian optimizer (e.g., using `geomstats` or `pymanopt`) to find the shortest path, then measure its length; or
  - *Inherited geometry*: For hypersphere $\mathbb{S}^{d-1}$ (VIDAR), use $d_g(z, z') = \arccos\langle z, z'\rangle$ directly.
- **Path sampling.** Sample paths that reflect operational usage:
  - HERALD: mutation trajectories with $\leq L_\star$ substitutions in validated epitope positions;
  - VIDAR: contiguous frame windows of length T with cumulative arc length $\leq L_\star$.
- **Slice-wise audits.** Compute $\hat{\varepsilon}(L_\star)$ separately for each slice (e.g., by age group, lighting condition, or generator). Flag slices where $\hat{\varepsilon} > \varepsilon_{\text{target}}$ for elevated E_{geom} or mandatory abstention.

6.2.2 Decision Rules and Gates

Table 8: **Distortion audit decision rules.** These thresholds are illustrative; actual values depend on domain and risk tolerance.

Condition	Interpretation	Action
$\hat{\varepsilon}(L_\star) < 0.10$	Low distortion, geometry reliable	Proceed; use geometry-based features (F1–F4 in VIDAR)
$0.10 \le \hat{\varepsilon}(L_\star) < 0.20$	Moderate distortion	Proceed with caution; down-weight geometric features or widen uncertainty
$\hat{\varepsilon}(L_\star) \ge 0.20$	High distortion	**Abstain** or suppress geometry; rely on auxiliary detectors only
$\lvert\rho_j - 1\rvert > 0.30$ for $>$ 10% of paths	Frequent outliers	Flag slice for investigation; may indicate OOD inputs or tracking failures

Example (VIDAR on deepfake validation data). Suppose Algorithm **??** yields:

$$\bar{\rho} = 1.03, \quad \sigma_\rho = 0.08,$$
$$\hat{\varepsilon}(L_\star) = \max_j |\rho_j - 1| = 0.18 \quad (\text{or } 95^{\text{th}} \text{ percentile: } 0.14).$$

This suggests moderate distortion. If using the 95^{th} percentile definition, $\hat{\varepsilon} = 0.14 < 0.20$ and we proceed; if using the max, $\hat{\varepsilon} = 0.18$ is borderline and we down-weight RIM features or increase abstention thresholds.

6.3 Phase 2: Margin Estimation and Non-Vacuity

The configuration margins $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$ from Definition **??** must be empirically validated to ensure that the non-vacuity condition

$$\kappa_{\text{soft}} - 2R\,\varepsilon(L_\star) \;>\; 0$$

holds on $\mathcal{D}_{\text{val}}$, where R bounds the geodesic radius of the operational region.

6.3.1 Protocol: Measuring Configuration Separation

Algorithm 2 Margin Estimation on Validation Pairs

Require: Encoder f_θ, validation set $\mathcal{D}_{\text{val}}$ with coarse labels $Y \in \{0, 1\}$
Ensure: Estimates $(\hat{\kappa}_{\text{hard}}, \hat{\kappa}_{\text{soft}})$ with confidence intervals
Sample N_{same} same-configuration pairs (x, x^+) with $h(c(f_\theta(x))) = h(c(f_\theta(x^+)))$
Sample N_{diff} different-configuration pairs (x, x^-) with $h(c(f_\theta(x))) \neq h(c(f_\theta(x^-)))$
Compute distances:
  Within-configuration: $\{d_{g_i^{(0)}}\}_{i=1}^{N_{\text{same}}}$ for same-config pairs
  Across-configuration: $\{d_{g,i}^{(1)}\}_{i=1}^{N_{\text{diff}}}$ for different-config pairs
Estimate margins:
  $\hat{\mu}_0 = \text{mean}(d_{g,i}^{(0)}), \quad \hat{\mu}_1 = \text{mean}(d_{g,i}^{(1)})$
  $\hat{\Delta}_g = \hat{\mu}_1 - \hat{\mu}_0$
  $\hat{\kappa}_{\text{hard}} = \hat{\mu}_0 + 2\hat{\sigma}_0$ ▷ Conservative: mean + 2 std of within-config
  $\hat{\kappa}_{\text{soft}} = \hat{\mu}_1 - 2\hat{\sigma}_1$ ▷ Conservative: mean - 2 std of across-config
Compute bootstrap CIs for $(\hat{\kappa}_{\text{hard}}, \hat{\kappa}_{\text{soft}})$ via resampling
Non-vacuity check: Verify $\hat{\kappa}_{\text{soft}} - 2R\,\hat{\varepsilon}(L_\star) > 0$ using $\hat{\varepsilon}$ from Phase 1
return $(\hat{\kappa}_{\text{hard}}, \hat{\kappa}_{\text{soft}})$, CIs, non-vacuity flag

Interpretation. If $\hat{\kappa}_{\text{hard}} < \hat{\kappa}_{\text{soft}}$ and the non-vacuity condition holds, the system has a validated ambiguity band $[\hat{\kappa}_{\text{hard}}, \hat{\kappa}_{\text{soft}}]$ where abstention should trigger. If $\hat{\kappa}_{\text{hard}} \geq \hat{\kappa}_{\text{soft}}$ or non-vacuity fails, the configuration structure is not sufficiently separated for reliable detection at horizon $L_\star$; either reduce $L_\star$, increase λ (smoothness), or collect more contrastive training data.

6.4 Phase 3: Error Budget Estimation

The compositional guarantee in Theorem **??** decomposes correctness into four error terms plus an independence slack. This subsection provides estimation protocols for each.

6.4.1 Geometry Error (E_{geom})

Definition. E_{geom} is the probability that the Davis manifold geometry behaves outside its validated regime in a way that materially corrupts the detection statistic S.

Operational proxy. On $\mathcal{D}_{\text{test}}$, flag samples as geometry failures if:

(a) Any path γ used to compute features has distortion $|\rho_\gamma - 1| > \varepsilon_{\max}$ (e.g., $\varepsilon_{\max} = 0.25$);

(b) Smoothness checks fail (e.g., large second-order finite differences in embeddings suggest high local curvature); or

(c) Paths exit the validated region (e.g., geodesic distance to training support exceeds a Mahalanobis threshold).

Estimator.

$$\hat{E}_{\text{geom}} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}\{\text{sample } i \text{ flagged as geometry failure}\}.$$

Compute separately on slices and worst-case across slices for a conservative bound.

6.4.2 Linkage Error (E_{link})

Definition. E_{link} is the probability that features fail to track geodesic distance even when geometry is good—e.g., monotonicity breaks, PIT nulls are misspecified, or feature fusion produces a statistic S uncorrelated with true configuration change.

Operational proxy. On $\mathcal{D}_{\text{test}}$, restrict to samples passing geometry checks (not flagged in Phase 1). For each, compute:

(a) Ground-truth configuration change Δc (binary: same vs. different);

(b) Predicted statistic S and its sign relative to threshold $s_\star$.

Flag as linkage failures any samples where S and Δc are anti-correlated (e.g., $S < s_\star$ but $\Delta c = 1$, or vice versa) despite passing geometry gates.

Estimator.

$$\hat{E}_{\text{link}} \;=\; \frac{\#\{\text{samples with anti-correlated } (S, \Delta c)\}}{\#\{\text{samples passing geometry checks}\}}.$$

6.4.3 Calibration Error (ξ)

Definition. ξ quantifies miscalibration in the mapping from the raw statistic S to the calibrated probability $\hat{p} = \text{calib}(S)$ in the high-risk region ($S > s_\star$).

Protocol: Expected Calibration Error (ECE).

1. Fit a calibrator (e.g., isotonic regression, Platt scaling, or temperature scaling) on $\mathcal{D}_{\text{cal}}$ to map $S \to \hat{p}$.
2. On $\mathcal{D}_{\text{test}}$, bin samples by $\hat{p}$ (e.g., 10 bins from 0 to 1).
3. For each bin B_k in the high-risk region ($\hat{p} > 0.5$ or $S > s_\star$), compute:

$$\text{ECE}_k \;=\; \big|\text{avg. predicted } \hat{p} \text{ in bin} - \text{fraction of true positives in bin}\big|.$$

4. Set $\hat{\xi} = \max_k \text{ECE}_k$ over high-risk bins, or use a weighted average.

Alternative: Brier score decomposition. Decompose the Brier score restricted to $S > s_\star$ into calibration and refinement components; use the calibration component as $\hat{\xi}$.

6.4.4 Abstention Failure (ζ)

Definition. ζ is the probability that the system fails to abstain when it should—i.e., issues a non-abstaining prediction (‘`info`’, ‘`warn`’, or ‘`high_risk`’) despite being out-of-regime.

Operational proxy. Define an *oracle abstention set* $\mathcal{A}_{\text{oracle}}$ on $\mathcal{D}_{\text{test}}$ as samples that fail any of:

- Distortion gate ($|\rho| - 1 > \varepsilon_{\max}$);
- Margin check ($d_g \in [\kappa_{\text{hard}}, \kappa_{\text{soft}}]$, i.e., in ambiguous band);
- Feature range check (S or individual features outside $[F_{\min}, F_{\max}]$ from training);
- Support check (e.g., Mahalanobis distance to training manifold exceeds threshold).

Let $\mathcal{A}_{\text{pred}}$ be the set of samples for which the Davis system actually abstains. Then:

$$\hat{\zeta} \;=\; \frac{\#(\mathcal{A}_{\text{oracle}} \setminus \mathcal{A}_{\text{pred}})}{\#\mathcal{A}_{\text{oracle}}} \;=\; \frac{\text{should-abstain but didn't}}{\text{total should-abstain}}.$$

6.4.5 Independence Slack (δ_{indep})

Definition. δ_{indep} captures the excess probability that *some* error occurs beyond what would be predicted by assuming $(E_{\text{geom}}, E_{\text{link}}, \xi, \zeta)$ are independent.

Estimator. On $\mathcal{D}_{\text{test}}$, flag each sample for each error type. Compute:

$$\hat{\varepsilon}_{\text{joint}} = \frac{\#\{\text{samples with } \geq 1 \text{ error}\}}{N_{\text{test}}},$$
$$\hat{\varepsilon}_{\text{mult}} = 1 - (1 - \hat{E}_{\text{geom}})(1 - \hat{E}_{\text{link}})(1 - \hat{\xi})(1 - \hat{\zeta}).$$

Then:

$$\hat{\delta}_{\text{indep}} \;=\; \max\{0,\, \hat{\varepsilon}_{\text{joint}} - \hat{\varepsilon}_{\text{mult}}\}.$$

If $\hat{\delta}_{\text{indep}} > 0.1$, the multiplicative bound is optimistic and a union-bound fallback should be used.

6.5 Phase 4: Abstention and Quality Gates

Abstention in Davis systems is not a post-hoc confidence threshold; it is a first-class outcome triggered when assumptions fail. This subsection formalizes abstention policies.

6.5.1 Multi-Stage Abstention Architecture

Table 9: **Abstention gates and decision logic.** Gates are applied sequentially; failure at any stage triggers '`insufficient_signal`'.

Gate	Condition	Interpretation
G1: Quality	Basic quality checks (resolution, blur, completeness)	Filters obviously broken inputs
G2: OOD	Mahalanobis distance, reconstruction error, or novelty score	Detects inputs far from training support
G3: Distortion	$\lvert\rho_\gamma - 1\rvert \le \varepsilon_{\max}$ for all paths γ	Ensures bounded-distortion regime
G4: Margin	$d_g \notin [\kappa_{\text{hard}}, \kappa_{\text{soft}}]$	Avoids ambiguous band
G5: Feature range	All features $F_k \in [F^{(k)}_{\min}, F^{(k)}_{\max}]$	Checks features are in-range
G6: Vacuity	$\hat{E}_{\text{geom}} + \hat{E}_{\text{link}} + \hat{\xi} + \hat{\zeta} < \tau_{\text{vac}}$	Global error budget below vacuity threshold

Decision tree.

1. If any of G1–G5 fail for a sample, output '`insufficient_signal`'.
2. If G6 fails (error budget exceeds $\tau_{\text{vac}} \approx 0.7$), downgrade all alerts or abstain globally.
3. Otherwise, compute S and proceed to threshold-based decisions: '`info`', '`warn`', or '`high_risk`'.

6.6 Phase 5: Calibration Methods

Calibration maps the raw detection statistic S to a probability $\hat{p} \approx \mathbb{P}(Y = 1 \mid S)$. This subsection describes three common approaches.

6.6.1 Isotonic Regression (Non-Parametric)

Fit a piecewise-constant, monotone map $S \mapsto \hat{p}$ via isotonic regression on $\mathcal{D}_{\text{cal}}$. This preserves the ranking of S while adjusting probabilities to match empirical frequencies.

Advantages: Flexible, no parametric assumptions.

Disadvantages: Can overfit on small calibration sets; requires careful binning.

6.6.2 Platt Scaling (Logistic)

Fit a logistic model:

$$\hat{p}(S) = \frac{1}{1 + \exp(-aS - b)},$$

where (a, b) are learned on $\mathcal{D}_{\text{cal}}$ via maximum likelihood.

Advantages: Simple, interpretable, works well when S is approximately linear in log-odds.

Disadvantages: Assumes a specific functional form.

6.6.3 Temperature Scaling (Neural)

If S is the output of a neural network before softmax, add a learned temperature $T > 0$:

$$\hat{p}(S) = \frac{\exp(S/T)}{\exp(S/T) + 1}.$$

Fit T on $\mathcal{D}_{\text{cal}}$ to minimize negative log-likelihood.

Recommendation. Use isotonic regression by default for Davis systems; fall back to Platt or temperature scaling if the calibration set is small (< 1000 samples).

6.7 Validation Checklists and Monitoring

Pre-Deployment Checklist

- ☐ Distortion audit complete: $\hat{\varepsilon}(L_\star) < \varepsilon_{\text{target}}$ on all validated slices
- ☐ Margins verified: $\hat{\kappa}_{\text{soft}} - 2R\hat{\varepsilon}(L_\star) > 0$ with CIs
- ☐ Error budget estimated: $(\hat{E}_{\text{geom}}, \hat{E}_{\text{link}}, \hat{\xi}, \hat{\zeta}, \hat{\delta}_{\text{indep}})$ documented
- ☐ Vacuity check: $\hat{E}_{\text{geom}} + \hat{E}_{\text{link}} + \hat{\xi} + \hat{\zeta} < \tau_{\text{vac}}$
- ☐ Calibration validated: ECE < 0.05 in high-risk region on $\mathcal{D}_{\text{test}}$
- ☐ Abstention gates tested: $\hat{\zeta} < 0.05$ on oracle abstention set
- ☐ Slice-wise audits: No slice has $\hat{\varepsilon} > 1.5 \times \varepsilon_{\text{target}}$ or $\hat{E}_{\text{sum}} > \tau_{\text{vac}}$
- ☐ Davis dossier prepared: All thresholds, data splits, and procedures version-controlled

6.7.1 Deployment Monitoring

Once deployed, Davis systems require continuous monitoring to detect distributional drift. Recommended metrics (computed on rolling windows, e.g., weekly):

Table 10: **Deployment monitoring dashboard (recommended metrics).**

Metric	**Trigger Condition for Re-Audit**
Abstention rate	$> 20\%$ increase vs. validation baseline
Distortion $\hat{\varepsilon}(L)$	Mean or 95$^{\text{th}}$ percentile exceeds $1.2 \times \varepsilon_{\text{target}}$
Feature statistics	Drift in $\mathbb{E}[F_k]$ or $\text{Var}(F_k)$ beyond 2 std from validation
Calibration drift	ECE increases by > 0.03 in high-risk region
Slice imbalance	Any slice has abstention rate $> 2\times$ global rate
Error budget	$\hat{E}_{\text{sum}}$ exceeds $0.8 \times \tau_{\text{vac}}$ (warning) or τ_{vac} (halt)

Response to drift. When monitoring triggers indicate assumption violations:

(a) **Minor drift**: Widen uncertainty, increase abstention thresholds, flag affected slices.

(b) **Moderate drift**: Pause high-risk alerts; rerun Phases 1–5 on recent data; update dossier.

(c) **Severe drift**: Halt deployment; retrain encoder with updated λ or path family; restart validation from Phase 1.

6.8 Computational Considerations

Davis systems trade off geometric fidelity for computational cost. This subsection quantifies typical overheads.

6.8.1 Training-Time Costs

Table 11: **Computational overhead of Davis construction (training).** Relative to baseline contrastive training without smoothness regularization.

Component	**Overhead Factor**	**Notes**
Jacobian regularization	1.5×–3×	Forward and backward for $\nabla_x f_\theta$
Laplacian regularization	2×–5×	Requires graph construction + sparse linear algebra
Distortion audits (validation)	0.1×–0.3× per epoch	Geodesic integration on sampled paths

Mitigation strategies.

- Use stochastic Jacobian estimation (random projections) instead of full Hessian.
- Run distortion audits asynchronously during training; cache geodesic computations.
- For frozen backbones (inherited margin), smoothness regularization is unnecessary.

6.8.2 Inference-Time Costs

Table 12: **Runtime overhead of Davis inference.** Per-sample latency relative to a single forward pass.

Operation	**Overhead Factor**	**Notes**
Embedding (f_θ)	1×	Baseline
Path features (F1–F4)	0.5×–1×	Geodesic differences, tangent projections
Distortion check	0.2×–0.5×	If geodesic solver used; negligible for hypersphere
PIT normalization	0.1×	ECDF lookups
Calibration map	0.05×	Isotonic or logistic eval
Total	**2×–3×**	Acceptable for offline/forensic use

Optimization. For real-time applications:

- Precompute null CDFs for PIT offline.
- Use closed-form geodesics (e.g., $\mathbb{S}^{d-1}$) to avoid iterative solvers.
- Cache embeddings for multi-frame or temporal inputs.

6.9 Worked Example: VIDAR Validation Protocol

To make the protocols concrete, we sketch how VIDAR (deepfake detection via identity dynamics) instantiates Phases 1–5.

Phase 1: Distortion audit.

1. Sample $N_{\text{path}} = 500$ short talking-head clips (real and synthetic) from $\mathcal{D}_{\text{val}}$.
2. Extract ArcFace embeddings $\{f_\theta(x_t)\}$ for frames $t = 1, \ldots, T$.
3. Normalize to $\mathbb{S}^{d-1}$; compute geodesic arc length $\ell_g = \sum_t \arccos\langle f_t, f_{t+1}\rangle$.
4. Compute Euclidean chord length $\ell_\delta = \sum_t \|f_t - f_{t+1}\|_2$.
5. Measure $\rho_j = \ell_\delta/\ell_g$ for each clip j; compute $\hat{\varepsilon} = 95^{\text{th}}$ percentile of $|\rho_j - 1|$.
6. Result: $\hat{\varepsilon}(L_\star = 0.3 \text{ rad}) = 0.12$ (acceptable; proceed).

Phase 2: Margin estimation.

1. Sample pairs: same-identity clips from $\mathcal{D}_{\text{val}}$ (real), different-identity clips (real vs. synthetic or different people).
2. Compute geodesic distances d_g between mean embeddings per clip.
3. Fit: $\hat{\kappa}_{\text{hard}} = 0.08$ rad, $\hat{\kappa}_{\text{soft}} = 0.18$ rad.
4. Non-vacuity: $0.18 - 2(0.5)(0.12) = 0.18 - 0.12 = 0.06 > 0$.

Phase 3: Error budget.

- $\hat{E}_{\text{geom}} = 0.08$ (8% of clips have $|\rho| > 0.20$; flagged)
- $\hat{E}_{\text{link}} = 0.05$ (5% anti-correlation between RIM features and labels)
- $\hat{\xi} = 0.04$ (ECE in high-risk region)
- $\hat{\zeta} = 0.03$ (3% fail to abstain when they should)
- $\hat{\delta}_{\text{indep}} = 0.02$ (joint vs. multiplicative slack)

Sum: $0.08 + 0.05 + 0.04 + 0.03 = 0.20 < \tau_{\text{vac}} = 0.70$.

Phase 4: Abstention gates. Configure gates G1–G5 with thresholds: quality (blur < 0.3), OOD (Mahalanobis < 3.0), distortion ($|\rho| < 0.20$), margin ($d_g \notin [0.08, 0.18]$), feature range (all RIM features in $[-3, 3]$ after PIT).

Figure 3: Davis roadmap (conceptual). *Top:* construction phase—contrastive and smoothness-regularized training produces a pullback manifold $(\mathcal{M}, g)$, a benign path family $\mathcal{P}(L_\star)$, and configuration structure with hard/soft margins and an ambiguity band. *Middle:* traversal phase—runtime observations x are embedded, paths are formed, features are extracted and PIT-normalized, and a scalar statistic S is computed with abstention gates. *Bottom:* theory—Theorem **??** ensures the existence of a Davis manifold under InfoNCE + smoothness; Theorem **??** provides a finite-variance detection bound with a compositional error budget; Section **??** discusses how to select an operational path horizon $L_\star$. HERALD and VIDAR (Tables **??** and **??**) instantiate this pipeline in virology and video forensics.

Phase 5: Calibration. Fit isotonic regression on $\mathcal{D}_{\text{cal}}$ mapping $S \to \hat{p}$; validate ECE $= 0.035 < 0.05$.

Outcome. VIDAR passes all validation checks and can deploy with the stated error budget. During deployment, monitor $\hat{\varepsilon}$, abstention rate, and ECE weekly; retrain if any drift beyond thresholds in Table **??**.

6.10 Summary and Transition to Discussion

This section has translated the theoretical Davis framework into actionable protocols:

- Algorithm **??** and **??** validate geometry and configuration structure.
- Subsection **??** provides estimation recipes for each error-budget component.
- Tables **??**, **??**, and **??** give decision rules and monitoring triggers.
- The VIDAR example (Subsection **??**) demonstrates end-to-end instantiation.

These protocols are not optional hygiene; they are the *enabling condition* for Theorem **??**. Without empirical validation of bounded distortion, non-vacuous margins, and sub-threshold error budgets, the compositional lower bound is vacuous and Davis guarantees do not apply. Section **??** now examines when these conditions hold in practice, discusses failure modes, and outlines the broader research agenda for geometry-first detection systems.

7 Discussion and Outlook

Figure **??** summarizes the Davis construction–traversal–guarantee pipeline: we (1) learn a Riemannian manifold and path family on which geodesic distance has semantic meaning; (2) traverse that geometry at runtime using Euclidean surrogates within a bounded-distortion regime; and (3) attach finite-variance, compositional error bounds and abstention policies that make the resulting system falsifiable.

7.1 Why Geometry?

Davis systems are built on a geometry-first view: we posit that the right notion of "distance" for a given domain lives on a curved functional state space, not in the raw observation coordinates. This choice is motivated by four recurring observations across domains.

Identity-preserving dynamics live on low-curvature manifolds. In HERALD, sequences that differ by a small number of epitope mutations induce nearby points and short geodesics on an antigenic manifold. In VIDAR, embeddings from a face-recognition backbone lie on a hypersphere, and clips induce identity trajectories with constrained velocity and curvature. In both cases, real-world dynamics respect continuity constraints that make a Riemannian description natural.

Bounded distortion makes Euclidean computations meaningful. Section **??** shows that contrastive training plus smoothness regularization yield a pullback metric g whose geodesics are well-approximated by straight chords over path families $\mathcal{P}(L_\star)$, with distortion controlled by $\varepsilon(L_\star)$. This lets us work with Euclidean distances and tangent-space statistics in practice while retaining a geometric interpretation: we know *where* those approximations are valid, and by how much they may distort.

Trajectories expose failures earlier than snapshots. High-risk events in Davis systems are typically *path properties*: a variant crossing an antigenic margin, an identity trajectory leaving a plausible band, a control policy departing a safe region. Representing observations as paths on $(\mathcal{M}, g)$ makes it natural to talk about velocity, curvature, and subspace leakage—and to detect abnormal behavior even when individual observations look unremarkable.

Geometry is where assumptions live. The bounded-distortion regime, margin non-vacuity $\kappa_{\text{soft}} - 2R\varepsilon(L_\star) > 0$, and path-family constraints are all geometric assumptions. By stating them explicitly on $(\mathcal{M}, g)$ and auditing them via distortion checks and margin measurements, we get falsifiable guarantees: when the geometry drifts or the path regime is violated, we know the theory no longer applies and must abstain or retrain.

7.2 When the Theory Applies—and When to Abstain

Section **??** (Theorem **??**) and Section **??** (Theorem **??**) are deliberately conditional. In practice, the Davis guarantees are intended to apply on a validated region $\mathcal{R}_{\text{valid}} \subseteq \mathcal{X}$ of in-distribution inputs where four conditions hold.

(T1) Geometry is stable. Distortion audits (Section **??**) show that along benign paths in $\mathcal{P}(L_\star)$, the ratio δ/d_g concentrates near 1 with tails bounded by a pre-registered $\varepsilon_\star$, and that temporal smoothness assumptions used in feature design are empirically satisfied.

(T2) Configurations are separated. Validation data support a positive gap between configurations that coarse labels h map to 0 vs. 1, after accounting for distortion: $\kappa_{\text{soft}} - 2R\varepsilon(L_\star) > 0$. In HERALD this appears as separation between vaccine-like and escape-like clusters in antigenic space; in VIDAR as a gap between same-identity and different-identity arcs on the hypersphere.

(T3) Features track geometry. The path features and aggregated statistic S satisfy a monotone linkage $\Delta_S \geq g(\Delta_g)$ on an empirically verified operational range, with non-trivial separation $\Delta_S > 0$ and finite second moments. Outside that range, the system either treats features as OOD or down-weights their contribution.

(T4) Calibration and abstention behave as specified. Calibration error in the high-risk region is bounded by ξ, and abstention failure by ζ, with both terms estimated under the protocols of

Section **??**. The error budget $E_{\text{geom}} + E_{\text{link}} + \xi + \zeta$ stays below a pre-registered vacuity threshold τ_{vac} on the deployment distribution.

Design principle. When empirical audits suggest that (T1)–(T4) may not hold on a slice or deployment—e.g., distortion tails are heavy, margins collapse, or calibration drifts—a Davis system should *increase abstention* rather than lowering thresholds to maintain apparent coverage.

Outside $\mathcal{R}_{\text{valid}}$ the finite-variance Cantelli bound may be loose and the compositional guarantee in Theorem **??** may become vacuous. The intended response is not to interpret the bound optimistically, but to treat such regimes as unvalidated: widen uncertainty, rely on external checks, or defer to human or laboratory adjudication.

7.3 Limitations and Failure Modes

Davis systems have several structural limitations.

Locality of the bounded-distortion regime. The distortion bounds in Section **??** are local in path length: they apply along paths in $\mathcal{P}(L)$ with $L \leq L_\star$ and bounded curvature. Recombination events, hard scene cuts, sudden lighting changes, or control discontinuities can produce large geodesic jumps where linearization fails. In such regimes the distinction between geodesic and Euclidean distance matters, and Davis guarantees should not be applied.

Dependence on the backbone. The geometry is induced by a particular encoder f_θ. If f_θ is biased or brittle—for example, face embeddings that compress distances for underrepresented groups, or sequence embeddings that collapse in novel lineages—then the Davis manifold inherits those failures. Conditions on the inherited margin (e.g., Assumption A5') must be checked slice-wise, not assumed.

Heavy-tailed noise and finite-variance bounds. The use of Cantelli's inequality in Theorem **??** buys generality (finite second moments suffice) at the cost of slower tail decay. In strongly heavy-tailed settings, the resulting bounds may be numerically weak. Sub-exponential or sub-Weibull refinements are an open direction (Section **??**).

Correlated failures. The multiplicative form of the bound assumes approximate independence among geometry, linkage, calibration, and abstention errors. When these error sources share common causes—for example, high curvature simultaneously degrading distortion and feature monotonicity, or OOD inputs breaking geometry and calibration together—the independence slack δ_{indep} can be large and the multiplicative bound optimistic; the union bound fallback becomes appropriate but may be vacuous.

Computational cost. Smoothness regularization, Jacobian penalties, and geodesic approximations can be substantially more expensive than training a black-box encoder with a simple loss. Runtime monitoring and distortion audits add overhead. For offline or forensic settings this cost may be acceptable; for real-time, high-throughput applications it can be a limiting factor.

Red flags. Empirically, the following are warning signs that a Davis instantiation is out of its validated regime: (i) distortion audits show large $\varepsilon(L_\star)$ or heavy tails, but coverage remains high; (ii) measured separation Δ_S collapses on deployment data; (iii) the estimated error budget exceeds τ_{vac}; or (iv) slice analyses show severe disparities that cannot be explained by data scarcity alone.

7.4 Relation to Existing Frameworks

Davis manifolds touch several established areas: metric learning, Riemannian machine learning, conformal prediction and selective classification, and domain-specific geometric systems such as HERALD and VIDAR. Table **??** summarizes the relationship at a high level.

Framework	**Primary focus**	**What Davis manifolds add**
Metric / contrastive learning	Embeddings where similar pairs are close, dissimilar pairs are far	Explicit Riemannian structure, bounded-distortion path families, and soft/hard configuration margins with ambiguity bands.
Riemannian ML (SPD nets, hyperbolic embeddings, etc.)	Optimization and representation on fixed manifolds	Data-driven *construction* of the manifold via pullback metrics, plus deployment-time distortion audits and abstention policies tied to the geometry.
Conformal / selective prediction, OOD detection	Coverage guarantees, abstention, uncertainty quantification	A geometric substrate that specifies when abstention should trigger (e.g., distortion or margin violations), and a compositional error budget that includes geometry and feature linkage.
HERALD, VIDAR (domain-specific systems)	Viral surveillance and deepfake forensics with geometry-based detectors	A unifying abstraction: both are instances of Davis systems with problem-specific manifolds, path families, and statistics; the present work generalizes their theory and operational protocols.

Table 13: Relation between Davis manifolds and selected prior frameworks.

Davis manifolds can be seen as an interface layer: metric learning and Riemannian ML supply tools for constructing $(\mathcal{M}, g)$; conformal prediction and selective classification supply tools for abstention and coverage; and domain knowledge supplies task-specific path families, configuration maps, and features.

7.5 The Davis Universality Conjecture

A central motivating question is whether Davis manifolds capture a broad class of safety-relevant temporal problems.

> **Conjecture (Davis universality).** Any supervised learning problem with identity-preserving temporal structure can be recast as path classification on a Davis manifold, provided that:

(U1) there exists an appropriate notion of "identity" or functional state that evolves continuously along benign trajectories;

(U2) a contrastive signal (same/different or similar/dissimilar pairs) is available to learn or validate a separation margin;

(U3) smoothness can be incentivized architecturally or via regularization so that bounded-distortion path families exist; and

(U4) configuration changes of interest exhibit a positive geodesic jump (a margin $\kappa > 0$) in the learned geometry.

Under these conditions, there exists a Davis manifold and path family for which the detection task can be formulated as classifying paths that cross configuration margins.

This conjecture may not hold in full generality, but it provides a concrete research program: identify domains that satisfy (or fail) the conditions, characterize what identity and smoothness mean in each, and study how far the Davis construction can be pushed.

Examples likely to satisfy (U1)–(U4).

- Protein or viral evolution (identity = functional antigenic state; paths = mutation trajectories).
- Video-based identity or pose tracking (identity = person or object; paths = short clips).
- Robot or autonomous-vehicle telemetry (identity = system state; paths = control trajectories).
- Speaker or instrument recognition over time (identity = source; paths = audio segments).

Examples that may violate the conjecture.

- Domains with rapidly mixing Markov dynamics and no meaningful notion of identity (violating (U1)).
- Settings where contrastive labels are unavailable or untrustworthy (violating (U2)).
- Adversarial environments that routinely induce discontinuous jumps in the relevant state, or where the generator explicitly targets the distortion and margin regimes (violating (U3) and (U4)).

7.6 Open Problems

Several technical questions are left intentionally open.

Sharper smoothness–distortion bounds. Section **??** uses a curvature-style bound $\varepsilon(L) \leq C'K(\lambda)L$ derived from smoothness penalties. Tightening this relationship—for example, via more precise control of sectional curvature or metric derivatives under Jacobian regularization—would sharpen the trade-off between distortion, path horizon, and regularization strength.

Beyond finite-variance tails. Our main detection bound relies on Cantelli and finite second moments. In some domains, heavier tails may warrant sub-exponential or robust alternatives that still admit an interpretable decomposition and integrate cleanly into the error budget.

Learning the path family. We treat $\mathcal{P}(L_\star)$ as a designed object (e.g., epitope-restricted mutation paths, identity trajectories with bounded arc length). Learning path families from data—subject to smoothness and curvature constraints—could reduce reliance on hand-crafted structure while maintaining interpretability.

Joint Davis geometry across modalities. HERALD and VIDAR each operate in a single embedding space. Many applications are naturally multi-modal (e.g., audio–visual, sequence–structure–clinical). Constructing joint manifolds and path families, with cross-modal distortion and margin guarantees, is an open challenge.

Adaptive error budgets. Section **??** treats E_{geom}, E_{link}, ξ, ζ, and δ_{indep} as periodically re-estimated but essentially static over a deployment window. Designing online estimators and control charts that adapt these quantities in real time without introducing look-ahead bias is an important systems problem.

Benchmarks and community standards. The Davis framework suggests a style of evaluation centered on distortion audits, error budgets, and abstention behavior. Curating benchmark suites and reporting standards that exercise these aspects across domains—for example, Davis-style tasks for sequences, video, robotics, and audio—is a prerequisite for systematic comparison and progress.

7.7 Reproducibility, Dossiers, and Deployment

Finally, translating Davis theory into practice requires structured documentation. Building on the operational protocols in Section **??**, we recommend that any deployed Davis system maintain a *Davis dossier* containing at least:

(D1) a description of the manifold construction: encoder architecture, training objectives, regularization, and the resulting distortion audits (empirical $\varepsilon(L)$ curves);

(D2) the configuration map and margins: how c and h are defined, values of $(\kappa_{\text{hard}}, \kappa_{\text{soft}})$, and evidence for non-vacuity;

(D3) the path family and features: definition of $\mathcal{P}(L_\star)$, feature design, and empirical separation Δ_S with confidence intervals;

(D4) error-budget estimates $(E_{\text{geom}}, E_{\text{link}}, \xi, \zeta, \delta_{\text{indep}})$ with methodology and slice-wise breakdowns;

(D5) the resulting lower bound from Theorem **??** for the intended deployment context, together with a vacuity threshold and abstention policy; and

(D6) version identifiers (code, weights, data snapshots) and a change log so that bounds and failures can be traced to specific system states.

The Davis framework is not a panacea for all temporal detection problems, but it offers a common pattern: construct a geometry in which distances and paths encode semantics; traverse that geometry using bounded-distortion surrogates; and attach explicit, compositional error budgets and abstention rules. HERALD and VIDAR show that this pattern can span domains as distant as viral evolution and deepfake forensics. We hope that future work will both stress-test and extend this template, clarifying where Davis manifolds truly help, where they fail, and how they might form part of a broader toolkit for safe, interpretable, and auditable AI systems.

Appendix A. Proofs and Technical Details

A.1. Proof of Lemma ?? (InfoNCE margin ⇒ embedding separation)

For completeness we restate the setting. Let $(x, x^+, x_1^-, \dots, x_K^-)$ be drawn from a contrastive sampling distribution $\mathcal{D}_{\mathrm{NCE}}$ in which (x, x^+) share the same fine-grained identity/configuration and each x_k^- has a different identity/configuration. Let

$$s_\theta(x, x') \;=\; -\frac{\|f_\theta(x) - f_\theta(x')\|_2^2}{2\tau^2}$$

denote the similarity score induced by the embedding f_θ with temperature $\tau > 0$, and let

$$\delta_\theta(x, x') \;=\; \|f_\theta(x) - f_\theta(x')\|_2$$

be the corresponding Euclidean distance. The InfoNCE population risk is

$$L_{\mathrm{NCE}}(\theta) \;=\; \mathbb{E}_{(x,x^+,x_1^-,\dots,x_K^-)\sim\mathcal{D}_{\mathrm{NCE}}}\left[-\log \frac{\exp\{s_\theta(x,x^+)\}}{\exp\{s_\theta(x,x^+)\} + \sum_{k=1}^K \exp\{s_\theta(x,x_k^-)\}}\right].$$

Define the *effective score margin*

$$m_{\mathrm{eff}}(\theta) \;:=\; \mathbb{E}\Big[s_\theta(x,x^+) - \frac{1}{K}\sum_{k=1}^K s_\theta(x,x_k^-)\Big]$$

and the distance gap

$$\Delta_{\mathrm{emb}} \;:=\; \mathbb{E}\big[\delta_\theta(x,x^-)\,\big|\,Y_{\mathrm{diff}}\big] - \mathbb{E}\big[\delta_\theta(x,x^+)\,\big|\,Y_{\mathrm{same}}\big].$$

Step 1: InfoNCE risk controls the score margin. For a single tuple, write

$$\Delta_k \;:=\; s_\theta(x,x_k^-) - s_\theta(x,x^+), \qquad k = 1,\dots,K.$$

Since $1 + \sum_k e^{\Delta_k} \ge \sum_k e^{\Delta_k}$ and $\log\big(\sum_k e^{u_k}\big) \ge \log K + \frac{1}{K}\sum_k u_k$ (AM–GM plus monotonicity of log), we obtain

$$\log\Big(1 + \sum_{k=1}^K e^{\Delta_k}\Big) \;\ge\; \log K + \frac{1}{K}\sum_{k=1}^K \Delta_k.$$

Taking expectations over $\mathcal{D}_{\mathrm{NCE}}$ and using the definition of $m_{\mathrm{eff}}(\theta)$ gives

$$L_{\mathrm{NCE}}(\theta) \;=\; \mathbb{E}\Big[\log\big(1 + \textstyle\sum_k e^{\Delta_k}\big)\Big] \;\ge\; \log K - m_{\mathrm{eff}}(\theta),$$

so

$$m_{\mathrm{eff}}(\theta) \;\ge\; \log K - L_{\mathrm{NCE}}(\theta). \tag{12}$$

Step 2: Population vs. empirical risk. Let $\widehat{L}_{\mathrm{NCE}}(\theta)$ denote the empirical InfoNCE risk on n i.i.d. samples. Assume a standard uniform-convergence bound for the loss class: there exists a complexity term C and universal constant $c > 0$ such that, with probability at least $1 - \gamma$,

$$L_{\mathrm{NCE}}(\theta) \;\le\; \widehat{L}_{\mathrm{NCE}}(\theta) \;+\; c\sqrt{\frac{C + \log(1/\gamma)}{n}}\;.$$

Combining with (**??**) gives, for any empirical minimiser $\hat{\theta}$,

$$m_{\mathrm{eff}}(\hat{\theta}) \;\ge\; \log K - \widehat{L}_{\mathrm{NCE}}(\hat{\theta}) - c\sqrt{\frac{C + \log(1/\gamma)}{n}}. \tag{13}$$

Step 3: Lipschitz transfer from scores to distances. On the bounded operational domain $\delta \in [0, D]$ (the in-distribution region for identity/configuration pairs), the score $s(\delta) = -\delta^2/(2\tau^2)$ is L_{eff}-Lipschitz in δ with

$$L_{\mathrm{eff}} \;=\; \sup_{\delta \in [0,D]} |s'(\delta)| \;=\; \frac{D}{\tau^2}.$$

Thus, for any two distances $u, v \in [0, D]$,

$$|s(u) - s(v)| \;\le\; L_{\mathrm{eff}}\, |u - v|.$$

Apply this with $u = \delta_\theta(x, x^+)$ and $v = \frac{1}{K}\sum_k \delta_\theta(x, x_k^-)$, take expectations, and rearrange to obtain

$$\Delta_{\mathrm{emb}} \;\ge\; \frac{m_{\mathrm{eff}}(\theta)}{L_{\mathrm{eff}}}.$$

Combining with (**??**) yields the claimed margin-to-separation relation from Lemma **??**, up to the usual $O(\sqrt{C/n})$ generalization slack.

Step 4: Symmetric label noise. Under symmetric label noise with rate $\eta < \frac{1}{2}$ (assumption A4), the effective margin contracts by a factor $(1 - 2\eta)$; this simply rescales the right-hand side of (**??**) and yields the $O(\eta)$ term in the lemma statement.

A.2. Explicit constants in the margin corollary

On the bounded domain $\delta \in [0, D]$ we have

$$L_{\mathrm{eff}} \;=\; \frac{D}{\tau^2} \qquad \Longrightarrow \qquad \frac{1}{L_{\mathrm{eff}}} \;=\; \frac{\tau^2}{D}.$$

Writing $a := \tau^2/D$ and absorbing the uniform-convergence constant into $b := c/L_{\mathrm{eff}} = c\tau^2/D$, the inequality in Lemma **??** can be expressed in the simplified form used in Section **??**:

$$\Delta_{\mathrm{emb}} \;\ge\; a\,\log K \;-\; b\sqrt{\frac{C + \log(1/\gamma)}{n}} \;-\; O(\eta),$$

with a, b depending only on the score temperature τ, the operational diameter D, and the capacity term C.

A.3. Cantelli bound for Davis systems (details for Proposition ??)

We recall the quantities from Section **??**. Let S be the scalar detection statistic, $Y \in \{0, 1\}$ the class label ($0 =$ benign, $1 =$ event), and

$$\mu_y = \mathbb{E}[S \mid Y = y], \qquad \sigma_y^2 = \mathrm{Var}(S \mid Y = y), \qquad \Delta_S = \mu_1 - \mu_0 > 0.$$

Define the midpoint threshold

$$s_\star = \frac{\mu_0 + \mu_1}{2} = \mu_0 + \frac{\Delta_S}{2},$$

and the class priors $\pi_y = \mathbb{P}(Y = y)$.

Cantelli inequality. For a real-valued random variable X with mean μ and variance $\sigma^2 < \infty$, Cantelli's (one-sided Chebyshev) inequality states that for any $a > 0$,

$$\mathbb{P}(X - \mu \geq a) \;\leq\; \frac{\sigma^2}{\sigma^2 + a^2}.$$

Apply this to $S \mid Y = 0$ with $a = \Delta_S/2$ to obtain

$$\mathbb{P}(S > s_\star \mid Y = 0) = \mathbb{P}(S - \mu_0 \geq \Delta_S/2 \mid Y = 0) \;\leq\; c_0 := \frac{\sigma_0^2}{\sigma_0^2 + (\Delta_S/2)^2}.$$

Similarly for $S \mid Y = 1$,

$$\mathbb{P}(S \leq s_\star \mid Y = 1) = \mathbb{P}(\mu_1 - S \geq \Delta_S/2 \mid Y = 1) \;\leq\; c_1 := \frac{\sigma_1^2}{\sigma_1^2 + (\Delta_S/2)^2},$$

so that

$$\mathbb{P}(S > s_\star \mid Y = 1) \;\geq\; 1 - c_1.$$

Posterior correctness above the midpoint. The posterior probability of $Y = 1$ given that S exceeds $s_\star$ is

$$\mathbb{P}(Y = 1 \mid S > s_\star) = \frac{\pi_1\, \mathbb{P}(S > s_\star \mid Y = 1)}{\pi_0\, \mathbb{P}(S > s_\star \mid Y = 0) + \pi_1\, \mathbb{P}(S > s_\star \mid Y = 1)}.$$

Using $\mathbb{P}(S > s_\star \mid Y = 0) \leq c_0$ and $\mathbb{P}(S > s_\star \mid Y = 1) \geq 1 - c_1$ gives the population bound

$$\mathbb{P}(Y = 1 \mid S > s_\star) \;\geq\; \frac{\pi_1(1 - c_1)}{\pi_0 c_0 + \pi_1(1 - c_1)}. \tag{14}$$

Finite-sample estimation. In practice, μ_y, σ_y^2 and π_y are replaced by empirical estimates $\hat{\mu}_y$, $\hat{\sigma}_y^2$, $\hat{\pi}_y$ computed on a validation set. Standard concentration arguments (e.g., for sub-Exponential or finite-variance variables) imply that with high probability the plug-in version of (**??**) deviates from its population value by at most an $O(\sqrt{1/n})$ term. In Proposition **??** we bundle these effects into the single estimation slack ε_{est}, yielding the stated lower bound.

An equivalent "odds form" sometimes useful numerically is

$$\frac{\mathbb{P}(Y=1 \mid S>s_\star)}{\mathbb{P}(Y=0 \mid S>s_\star)} \;\geq\; \frac{\pi_1}{\pi_0}\,\frac{1-c_1}{c_0},$$

from which (**??**) follows by algebra.

A.4. Independence slack, union bound, and vacuity (supporting material for Theorem ??)

Recall the four component error probabilities from Section **??**:

$$E_{\text{geom}}, \quad E_{\text{link}}, \quad \xi, \quad \zeta,$$

corresponding to geometry failure, linkage/feature failure, calibration error and abstention failure, respectively. Write

$$G_{\text{geom}} = \{\text{geometry behaves as intended}\},$$

and similarly define G_{link}, G_{cal}, G_{abst} as the complements of the corresponding error events, so that

$$\mathbb{P}(G_{\text{geom}}^{c}) = E_{\text{geom}}, \quad \mathbb{P}(G_{\text{link}}^{c}) = E_{\text{link}}, \quad \mathbb{P}(G_{\text{cal}}^{c}) = \xi, \quad \mathbb{P}(G_{\text{abst}}^{c}) = \zeta.$$

Let

$$G \;=\; G_{\text{geom}} \cap G_{\text{link}} \cap G_{\text{cal}} \cap G_{\text{abst}}$$

denote the global "good" event on which all components behave.

Approximate independence and slack. If the four error events were independent, we would have

$$\mathbb{P}(G) = \prod_{i\in\{\text{geom,link,cal,abst}\}} (1-E_i) \;=:\; P_{\text{indep}}.$$

In practice, errors may be positively correlated (e.g., heavy OOD shifts that simultaneously break geometry and calibration). We therefore introduce a non-negative *independence slack* $\delta_{\text{indep}} \geq 0$ defined implicitly by

$$\mathbb{P}(G) \;\geq\; P_{\text{indep}} - \delta_{\text{indep}}. \tag{15}$$

When independence (or approximate independence) is empirically plausible, δ_{indep} is small; when errors co-occur frequently, δ_{indep} can be large and the product P_{indep} is overly optimistic.

In Section **??** we estimate δ_{indep} from validation data via

$$\hat{\varepsilon}_{\text{joint}} = \mathbb{P}\big(\text{at least one error occurs}\big), \quad \hat{\varepsilon}_{\text{mult}} = 1 - \prod_i (1-\hat{E}_i),$$

and set $\hat{\delta}_{\text{indep}} = \max\{0, \hat{\varepsilon}_{\text{joint}} - \hat{\varepsilon}_{\text{mult}}\}$; this ensures (**??**) holds with high probability up to sampling noise.

Union-bound fallback and vacuity. Even without any independence assumptions we always have the union bound

$$\mathbb{P}(G) = 1 - \mathbb{P}(G^{c}) \;\geq\; 1 - (E_{\mathrm{geom}} + E_{\mathrm{link}} + \xi + \zeta).$$

This bound can be used in place of (**??**) when empirical estimates suggest a large δ_{indep}. In regimes where the right-hand side becomes non-positive, the global bound in Theorem **??** is vacuous; we explicitly clip it at zero in that case and treat the system as operating outside its validated regime.

A.5. Proof of Theorem ?? (existence of Davis manifolds)

Proof. Theorem **??** asserts that, under assumptions D1–D3 in Section **??**, one can choose training hyperparameters and operational scales so that the resulting tuple satisfies Definition **??**. The argument is a direct combination of Lemma **??** and Proposition **??**.

Step 1: Configuration separation from InfoNCE. Assumption D1 states that contrastive training (or inherited margins) produces a positive distance gap between different configurations. Concretely, Lemma **??** gives a lower bound

$$\Delta_{\mathrm{cfg}} \;:=\; \mathbb{E}\big[\delta(x, x^{-}) \mid \text{different configuration}\big] - \mathbb{E}\big[\delta(x, x^{+}) \mid \text{same configuration}\big] \;>\; 0$$

up to the usual finite-sample and label-noise slack. Define hard and soft configuration margins by

$$\kappa_{\mathrm{hard}} := \frac{1}{4}\,\Delta_{\mathrm{cfg}}, \qquad \kappa_{\mathrm{soft}} := \frac{1}{2}\,\Delta_{\mathrm{cfg}}.$$

By construction, configurations whose representatives lie within κ_{hard} of a reference set are comfortably inside a class, while configurations separated by at least κ_{soft} cannot be confused by small perturbations. This matches the configuration structure of Definition **??**.

Step 2: Bounded distortion along benign paths. Assumption D2 posits that the construction objective includes a smoothness term with weight λ, and that the resulting metric satisfies the curvature control conditions of Proposition **??**. For any benign path family $\mathcal{P}(L)$ (Definition **??**), Proposition **??** then yields a distortion profile of the form

$$\varepsilon(L) \;\leq\; C' K(\lambda) L,$$

where $K(\lambda)$ is a curvature proxy that decreases as λ increases and $C' > 0$ is a geometric constant depending only on the architecture and data manifold.

Fix an operational path horizon $L_\star$ and a ball radius R as in assumption D3. Because $K(\lambda) \to 0$ as $\lambda \to \infty$, we can choose λ large enough that

$$\varepsilon(L_\star) \;\leq\; \frac{\kappa_{\mathrm{soft}}}{2R}.$$

Step 3: Non-vacuity of the configuration margin. With the choice above we have

$$\kappa_{\mathrm{soft}} - 2R\,\varepsilon(L_\star) \;\geq\; \kappa_{\mathrm{soft}} - 2R \cdot \frac{\kappa_{\mathrm{soft}}}{2R} \;=\; 0,$$

and in fact strict inequality holds once we back off slightly from the limiting value of λ. Thus the non-vacuity condition

$$\kappa_{\text{soft}} - 2R\,\varepsilon(L_\star) > 0$$

is satisfied, ensuring that configuration separation survives bounded-distortion perturbations along all paths in $\mathcal{P}(L_\star)$ of length at most R.

Step 4: Verification of the Davis-manifold definition. Collect the components

$$(\mathcal{M}_\theta, g_\theta, \mathcal{P}(L_\star), \varepsilon(\cdot), c, h, \kappa_{\text{hard}}, \kappa_{\text{soft}})$$

where $\mathcal{M}_\theta$ is the learned manifold with metric g_θ, $\mathcal{P}(L_\star)$ and $\varepsilon(\cdot)$ are as above, c and h are the configuration and coarse-label maps from Section **??**, and $\kappa_{\text{hard}}, \kappa_{\text{soft}}$ are defined in Step 1. By construction and the previous steps, this tuple satisfies all items in Definition **??**: configuration separation, bounded distortion along benign paths, and non-vacuous margins at the chosen horizon $L_\star$. The explicit parameter choices match those stated in Theorem **??**. □

A.6. Proof of Proposition ?? (geometric margin ⇒ statistic separation)

Proof. Proposition **??** formalizes the link between a geometric margin on the Davis manifold and separation in the scalar statistic S used in Theorem **??**.

Let $z_0, z_1 \in \mathcal{M}$ be representative points for two coarse configurations with $h(c(z_0)) = 0$ and $h(c(z_1)) = 1$. By the configuration-margin definition, there exist reference points $z^{(0)}, z^{(1)}$ such that

$$d_g(z_0, z^{(0)}) \le \kappa_{\text{hard}}, \qquad d_g(z_1, z^{(1)}) \ge \kappa_{\text{soft}},$$

and any path $\gamma \in \mathcal{P}(L_\star)$ that moves from the first configuration to the second must traverse at least a geodesic distance $\kappa_{\text{soft}} - \kappa_{\text{hard}}$ somewhere along its length. In particular,

$$d_g(z_0, z_1) \;\ge\; \kappa_{\text{soft}} - \kappa_{\text{hard}}.$$

By bounded distortion at the operational horizon (Section **??**), every such path satisfies

$$\delta(z_0, z_1) \;\ge\; (1 - \varepsilon_\star)\, d_g(z_0, z_1) \;\ge\; (1 - \varepsilon_\star)\,(\kappa_{\text{soft}} - \kappa_{\text{hard}}),$$

where $\varepsilon_\star := \varepsilon(L_\star)$.

Now let D denote the distance-like quantity used as the argument for the path features $\{\phi_k\}$ in Section **??** (for example, an average or maximum Euclidean distance along the path). By assumption, each feature ϕ_k is differentiable and monotonically increasing in D on the operational range $[d_{\min}, d_{\max}]$, with derivative

$$\phi_k'(d) \;\ge\; m_k > 0, \qquad d \in [d_{\min}, d_{\max}].$$

The scalar statistic takes the form

$$S \;=\; \sum_{k=1}^{K} w_k\, \phi_k(D_k),$$

with non-negative weights $w_k \geq 0$. For a benign configuration we write $D_k^{(0)}$ and $S^{(0)}$; for an event configuration $D_k^{(1)}$ and $S^{(1)}$.

Along the separating path considered above we have

$$D_k^{(1)} - D_k^{(0)} \ \gtrsim\ \delta(z_1, z_0) \ \geq\ (1 - \varepsilon_\star)\,(\kappa_{\text{soft}} - \kappa_{\text{hard}}),$$

up to constants that can be absorbed into the feature slopes; this uses the fact that the path features are Lipschitz in the underlying Euclidean distances, which in turn are bounded below by the distorted geodesic distance.

By the mean value theorem and monotonicity of each ϕ_k,

$$\phi_k(D_k^{(1)}) - \phi_k(D_k^{(0)}) \ \geq\ m_k\,(D_k^{(1)} - D_k^{(0)}).$$

Summing with weights $w_k \geq 0$ gives

$$S^{(1)} - S^{(0)} = \sum_k w_k \Big(\phi_k(D_k^{(1)}) - \phi_k(D_k^{(0)})\Big) \ \geq\ \Big(\sum_k w_k m_k\Big)\,(D^{(1)} - D^{(0)}),$$

where $D^{(y)}$ denotes a representative aggregate distance for class y (absorbing detector-specific constants into the m_k). Taking expectations over the class-conditional distributions yields

$$\Delta_S := \mu_1 - \mu_0 \ \geq\ \alpha\,(1 - \varepsilon_\star)\,(\kappa_{\text{soft}} - \kappa_{\text{hard}}),$$

with

$$\alpha := \sum_k w_k m_k > 0.$$

The non-vacuity condition $\kappa_{\text{soft}} > \kappa_{\text{hard}}$ and bounded distortion $\varepsilon_\star < 1$ imply $\Delta_S > 0$, as claimed. □

A.7. Proof of Theorem ?? (Main Davis bound)

Proof. We prove the lower bound on $\mathbb{P}(Y = 1 \mid H)$ by conditioning on the "good" event G in which all four error sources behave as intended.

Step 1: Conditional bound under G. Recall that

$$G = G_{\text{geom}} \cap G_{\text{link}} \cap G_{\text{cal}} \cap G_{\text{abst}},$$

where G_{geom} is the event that the bounded–distortion regime holds on $\mathcal{P}(L_\star)$, G_{link} that the features track geometry as intended, G_{cal} that calibration is adequate in the high–risk region, and G_{abst} that abstention fires whenever the construction assumptions fail. Under G, Proposition **??** yields a positive statistic gap $\Delta_S > 0$ with finite class–conditional variances, and all assumptions of Proposition **??** are satisfied at the operational threshold $s_\star$. Hence

$$\mathbb{P}(Y = 1 \mid S > s_\star, G) \ \geq\ \frac{\pi_1(1 - c_1)}{\pi_0 c_0 + \pi_1(1 - c_1)} \ -\ \varepsilon_{\text{Cantelli}}, \tag{16}$$

where c_0, c_1 are the Cantelli tails and $\varepsilon_{\text{Cantelli}}$ collects finite-sample estimation error in the moments. The event H ("high_risk" alert) is chosen so that $H \subseteq \{S > s_\star\}$, so

$$\mathbb{P}(Y = 1 \mid H, G) \;\geq\; \frac{\pi_1(1 - c_1)}{\pi_0 c_0 + \pi_1(1 - c_1)} \;-\; \varepsilon_{\text{Cantelli}}. \tag{17}$$

Step 2: Law of total probability over G and positive correlation. Decompose with respect to G and G^{c}:

$$\mathbb{P}(Y = 1 \mid H) = \mathbb{P}(Y = 1 \mid H, G)\,\mathbb{P}(G \mid H) + \mathbb{P}(Y = 1 \mid H, G^{\text{c}})\,\mathbb{P}(G^{\text{c}} \mid H).$$

Since $\mathbb{P}(Y = 1 \mid H, G^{\text{c}}) \geq 0$, we obtain the crude lower bound

$$\mathbb{P}(Y = 1 \mid H) \;\geq\; \mathbb{P}(Y = 1 \mid H, G)\,\mathbb{P}(G \mid H). \tag{18}$$

Using the bound from Step 1, we now relate the conditional probability $\mathbb{P}(G \mid H)$ to the prior error budget $\mathbb{P}(G)$. We make the *positive correlation assumption*:

$$\mathbb{P}(G \mid H) \;\geq\; \mathbb{P}(G). \tag{19}$$

This assumption states that an alert H is at least as likely to come from a well-behaved, in-regime system (event G) as from a broken one. Intuitively, broken systems may either fail to alert or produce spurious alerts, but a properly calibrated Davis system with working geometry and abstention is more likely to issue meaningful high-risk alerts than one where multiple components have failed. With this assumption, we arrive at

$$\mathbb{P}(Y = 1 \mid H) \;\geq\; \mathbb{P}(G)\left(\frac{\pi_1(1 - c_1)}{\pi_0 c_0 + \pi_1(1 - c_1)} \;-\; \varepsilon_{\text{Cantelli}}\right). \tag{20}$$

Step 3: Lower-bounding $\mathbb{P}(G)$ via the error budget. By definition,

$$\mathbb{P}(G^{\text{c}}_{\text{geom}}) = E_{\text{geom}}, \quad \mathbb{P}(G^{\text{c}}_{\text{link}}) = E_{\text{link}}, \quad \mathbb{P}(G^{\text{c}}_{\text{cal}}) = \xi, \quad \mathbb{P}(G^{\text{c}}_{\text{abst}}) = \zeta.$$

Under the approximate-independence modelling assumption and the dependence slack δ_{indep} defined in Appendix A.4, we have

$$\mathbb{P}(G) \;\geq\; (1 - E_{\text{geom}})(1 - E_{\text{link}})(1 - \xi)(1 - \zeta) \;-\; \delta_{\text{indep}}.$$

Substituting this into (**??**) and folding $\varepsilon_{\text{Cantelli}}$ together with the finite-sample errors in the estimates of $E_{\text{geom}}, E_{\text{link}}, \xi, \zeta$ and δ_{indep} into a single term ε_{est} yields

$$\mathbb{P}(Y = 1 \mid H) \;\geq\; \Big[(1 - E_{\text{geom}})(1 - E_{\text{link}})(1 - \xi)(1 - \zeta) - \delta_{\text{indep}}\Big]\frac{\pi_1(1 - c_1)}{\pi_0 c_0 + \pi_1(1 - c_1)} \;-\; \varepsilon_{\text{est}}.$$

This is exactly the lower bound stated in Theorem **??**. $\square$