Detecting Deepfakes via Hamiltonian Dynamics

Harry Cheng, *Member, IEEE*, Ming-Hui Liu, Tianyi Wang, *Member, IEEE*, Weili Guan, *Member, IEEE*, Liqiang Nie, *Senior Member, IEEE*, Mohan Kankanhalli, *Fellow, IEEE*

***Abstract*—Driven by the rapid development of generative AI models, deepfake detectors are compelled to undergo periodic recalibration to capture newly developed synthetic artifacts. To break this cycle, we propose a new perspective on deepfake detection: moving from static pattern recognition to dynamical stability analysis. Specifically, our approach is motivated by physics-inspired priors: we hypothesize that natural images, as products of dissipative physical processes, tend to settle near stable, low-energy equilibria. In contrast, generative models optimize for statistical similarity to real images but do not explicitly enforce structural constraints such as geometric smoothness, leaving deepfakes more likely to occupy unstable, high-energy states. To operationalize this, we introduce Hamiltonian Action Anomaly Detection (HAAD), comprising three contributions: i) We model the image latent manifold as a potential energy surface. Under this hypothesis, real images are expected to produce basin-like low-energy responses, whereas fake images are more likely to induce high-potential, high-gradient responses. ii) We employ Hamiltonian-inspired dynamics as a stability probe. By releasing latent states from rest, samples near stable regions remain bounded, while high-gradient samples produce larger trajectory responses. iii) We quantify these dynamic behaviors through two trajectory statistics, *i.e.*, Hamiltonian action and energy dissipation. Extensive experiments show that HAAD outperforms evaluated state-of-the-art baselines on challenging cross-dataset transfer benchmarks, supporting a physics-inspired stability prior for digital forensics.**

***Index Terms*—Deepfake Detection, AIGI, Hamiltonian Dynamics, Physics-Inspired Inductive Bias, Latent Stability Analysis.**

I. Introduction

The rapid advancement of generative models has blurred the boundary between reality and fabrication [1], [2]. While fueling creativity, these techniques have lowered the barrier for malicious exploitation, such as 'deepfakes' that threaten information ecosystems [3]. As a result, developing detection frameworks has been an urgent societal need.

The research community has proposed a number of detectors [4]–[8]. The prevailing paradigm focuses on identifying *specific static artifacts*, such as blending inconsistencies [9] or upsampling fingerprints [10]. However, these approaches suffer from a fundamental limitation: they tend to overfit to these artifacts [11]. As generative architectures evolve, characteristic artifacts change drastically [12], rendering existing detectors ineffective and trapping the field in a 'cat-and-mouse' cycle.

Harry Cheng, Tianyi Wang, and Mohan Kankanhalli are with the School of Computing, National University of Singapore, Singapore (e-mail: xaCheng1996@gmail.com; terry.ai.wang@gmail.com, dcsmsk@nus.edu.sg). Ming-Hui Liu is with the School of Software, Shandong University, China (e-mail: liuminghui@mail.sdu.edu.cn). Weili Guan is with the School of Information Science and Technology, Harbin Institute of Technology (Shenzhen), China (e-mail: honeyguan@gmail.com). Liqiang Nie is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China (e-mail: nieliqiang@gmail.com).

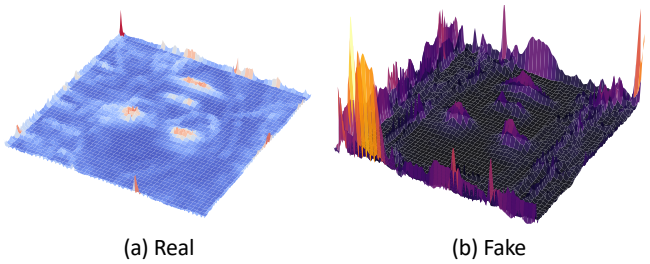

Fig. 1. Visualization of microscopic structural irregularities, measured by Laplacian magnitude and averaged over 1,000 images from the Celeb-DF++ dataset [13]. (a) Real images yield a smoother averaged manifold surface. (b) Forged images exhibit more jagged high-frequency spikes.

To break this cycle, we seek a detection cue from the *physics of image formation* rather than the specific artifacts of any particular generator, which remains valid as generative architectures evolve. Specifically, real images are formed by physical imaging processes, *i.e.*, they are captured by camera sensors through light reflected from surfaces, which inherently enforce local regularities such as geometric smoothness between neighboring pixels and photometric consistency of shading across surfaces. In contrast, fake images are produced by generative models (*e.g.*, GANs and diffusion models) trained with distribution-level objectives that contain no explicit penalty for violating such local structure. As a result, while generators can match natural images at the global statistical level, they often leave *microscopic structural irregularities* [14] that physically formed real images do not exhibit. As shown in Figure 1, real images form a smooth manifold surface, whereas deepfakes are punctuated by high-frequency jagged spikes.

To turn this observation into a quantitative signal, we draw an analogy with *physical systems*. Specifically, in such systems, configurations that are smooth and relaxed (*e.g.*, a flat elastic membrane) correspond to low-energy states, whereas locally stretched or jagged configurations store potential energy as elastic tension. That is, they reside in high-energy states. By this analogy, we interpret the microscopic structure of the images in Figure 1 as a form of *energy*, where images exhibiting greater structural inconsistency (*i.e.*, fake images) should correspond to higher energy values, while images satisfying the geometric and photometric regularities receive low energy. It is worth noting that this energy is defined by physical regularities rather than by the specific artifacts of any particular generator. Therefore, the resulting real-vs-fake energy gap is less tied to any single generator and is expected to generalize across evaluated non-adaptive transfer settings.

However, directly comparing this static energy between real and fake images can be unreliable, as the gap may be subtle. To amplify this gap, we shift our focus from static measurement

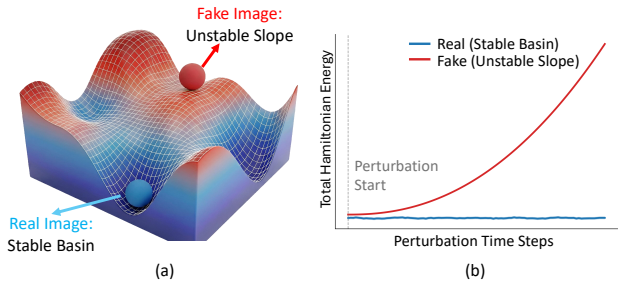

Fig. 2. (a) Manifold Stability Hypothesis. Real images are expected to concentrate near stable basins (blue), while deepfakes are more likely to induce high-gradient slopes (red). (b) Hamiltonian-inspired trajectories. Basin-like samples remain bounded under perturbation, while high-gradient samples exhibit larger energy divergence (high action).

to *dynamical evolution*. Specifically, we hypothesize that real images tend to reside in lower-energy states, while deepfakes are more likely to occupy higher-energy states. If we further conceptualize each image as a physical particle on an energy landscape, real particles will more often sit near stable basins, while fake particles will more often appear on steep slopes (as shown in Figure 2a). We term this assumption the **Manifold Stability Hypothesis** (MSH)[1]. To operationalize it, we model the particle's evolution within a **Hamiltonian mechanics** framework [15]. In this framework, the energy landscape exerts forces on the particle through its gradient, and we simulate the resulting trajectory. As illustrated in Figure 2b, particles near stable basins remain approximately stationary, whereas particles on steep slopes accelerate along the gradient. By tracking these trajectories, we convert subtle static energy differences into macroscopic motion patterns.

To turn this hypothesis into a trainable stability probe, we propose Hamiltonian Action Anomaly Detection (HAAD). Specifically, HAAD consists of three components: **i) Realizing the energy landscape** (*learnable potential* V). Under the MSH, the energy function should statistically separate basin-like real responses from high-gradient fake responses. We instantiate this landscape as a learnable potential V built from two physics-motivated priors: geometric smoothness (via graph Laplacians, penalizing abrupt feature transitions between neighboring patches) and photometric consistency (via Lambertian shading constraints, penalizing implausible illumination patterns). Images that violate these priors receive high potential energy, making them more likely to generate slope-like trajectory responses. **ii) Carrying out the dynamics** (*symplectic evolution*). We stipulate that each particle evolves under the force induced by V. To realize this, we employ a short-horizon symplectic integrator, which is designed to reduce solver-induced drift while probing the predicted behavior: particles near minima stay approximately put, while particles on slopes accelerate along the gradient. This can amplify small static energy gaps into larger trajectory discrepancies. **iii) Reading out the motion patterns** (*Action* S *and Dissipation* D). We distinguish real from fake by qualitative trajectory behavior via two statistics: *Action* S, the accumulated energy cost measuring how forcefully the particle is driven, and *Dissipation* D, the energy drift measuring deviation from stable evolution. Together, they serve as discriminative features for classification. Through these three closely associated components, HAAD anchors deepfake detection on a stability-based inductive bias rather than transient artifacts.

We evaluate HAAD on several deepfake benchmarks [16]–[19], where the cross-dataset and cross-generator experiments show strong transfer performance compared to several state-of-the-art (SoTA) baselines. Our contributions are three-fold:

- We introduce a physics-inspired stability prior for deepfake detection. We posit that real images tend to reside near low-energy basins while deepfakes are more likely to occupy high-energy slopes of a physics-motivated energy landscape. This reframes detection from static pattern matching to *dynamical stability analysis*.
- We propose *Hamiltonian Action Anomaly Detection*, which learns a potential V from physical priors, evolves image representations with a short-horizon symplectic integrator, and extracts two trajectory statistics as discriminative features for classification.
- Extensive experiments support the proposed MSH as an effective stability prior. HAAD achieves strong performance among SoTA baselines under cross-dataset and cross-generator experiments.

II. Related Work

A. Deepfake and Synthetic Image Detection

Our work addresses both deepfake manipulations (*e.g.*, face swapping) [3], [12], [20] and general image synthesis (*e.g.*, natural scenes) [2], [21], [22]. Across these domains, the research community is focused on generalization against unseen forgery methods. Existing approaches typically fall into two paradigms: i) **Artifact Mining** focuses on identifying explicit statistical anomalies left by generative processes [23], [24]. Researchers have exploited frequency discrepancies [25], [26], blending boundary traces [9], [27], and reconstruction errors [22], [28]. ii) **Foundation Model Adaptation** employs the rich representations of pre-trained vision-language models [29]–[31]. For instance, VLFFD [32] utilizes CLIP [33] to exploit semantic priors for detection. Effort [34] integrates Singular Value Decomposition (SVD) into pre-trained vision encoder to enhance sensitivity to structural artifacts. Although these methods achieve impressive performance, they treat the input image as a *static map*, classifying it based on learned features. This creates a 'lag': as generative models evolve to suppress these specific artifacts, detectors become obsolete. Some recent work goes beyond static appearance. For instance, FAMM [35] detects compressed deepfake videos by analyzing facial muscle motion patterns rather than individual frames. However, these motion-based approaches rely on temporal signals across video frames, which are unavailable for single-image deepfakes, and do not extend to general AIGI detection. Our work occupies a complementary niche: we introduce a *latent stability probe* based on Hamiltonian dynamics, which detects structural instability from a single image's feature representation without requiring temporal sequences or per-frame motion estimation.

[1]Please see Section III-A for details.

B. Physics-Informed Vision and Dynamics

Unlike purely data-driven approaches, physics-based vision explicitly models the structural and photometric constraints in natural image formation [36]. For instance, inverse rendering approaches demonstrate that real images adhere to photometric laws [37]. Parallel to this, physics-inspired neural networks incorporate physical laws as inductive biases for representation learning [14], and Hamiltonian mechanics has been integrated into machine learning to model continuous dynamical systems [15], [38]. Hamiltonian Neural Networks (HNNs) [39] enforce energy conservation to learn physical laws from observed trajectory data [40]. However, these methods are predominantly confined to *simulation* or *future forecasting* in closed environments. To the best of our knowledge, our HAAD is the first work to cast deepfake detection as *dynamical stability analysis* under a Hamiltonian framework. Instead of learning *static* physical consistency, our approach constructs a learnable potential energy surface based on physical priors and treats the image representation as a particle system. By subjecting the image features to a symplectic evolution, we transform the detection problem into a *dynamical stability test*, revealing the dynamical instability of synthesized images predicted by the Manifold Stability Hypothesis.

III. Methodology

Before presenting the HAAD architecture, we first formalize the theoretical framework motivating its design. Specifically, we cast deepfake detection as *dynamical stability analysis* on a learned potential energy landscape. Each latent feature is modeled as a dynamical state $(\mathbf{q}, \mathbf{p})$ that evolves under Hamiltonian mechanics, where $\mathbf{q}$ encodes image content and $\mathbf{p}$ is the canonical momentum driving the probe. Under this formulation, real images are hypothesized to reside near stable equilibria, since they are shaped by natural physical processes. By contrast, deepfakes, trained with objectives that do not explicitly enforce local structural regularities, tend to occupy unstable states on the landscape. We formalize this intuition as the *Manifold Stability Hypothesis* (Theorem III.2) and derive detection cues from its predicted dynamical behavior.

A. Theoretical Formulation

Directly verifying whether an image satisfies every physical regularity of natural scenes (surface smoothness, illumination coherence, geometric consistency, *etc.*) is infeasible: these regularities form an open-ended set, and any explicit enumeration risks missing cases by construction. Therefore, HAAD does not aim for such completeness. Instead, we construct a *potential energy landscape* based on a set of intuitive physical priors (detailed in Section III-B1), where a point's elevation measures the total extent to which it violates those priors.

Geometric Definitions of the Data Manifold. Let $E : \mathcal{X} \to \mathcal{Q} \subseteq \mathbb{R}^d$ be an encoder mapping the image space $\mathcal{X}$ to a latent manifold $\mathcal{Q}$. We denote $\mathbf{q} = E(\mathbf{x}) \in \mathcal{Q}$ as a latent position, where $\mathbf{x}$ is an input image. We interpret this manifold as a physical system governed by a scalar potential field.

Definition III.1 (**Latent Potential Energy**)**.** We define a potential function $V : \mathcal{Q} \to \mathbb{R}$ that measures how much a state $\mathbf{q} \in \mathcal{Q}$ deviates from the physics-motivated regularities of natural images. Conceptually, this aligns with Energy-Based Models (EBMs), where the probability density of real data follows a Gibbs distribution $p(\mathbf{q}) \propto \exp(-V(\mathbf{q}))$. Within this framework, the **negative** gradient field $-\nabla V(\mathbf{q})$ represents the *restorative force* driving the system toward high-density (low-energy) equilibrium regions.

Physical Motivation. Dissipative mechanics [41] and Lyapunov stability theory [42] motivate the intuition that physical systems tend to lose energy until they settle near stable, low-energy equilibria. The Principle of Least Action provides a complementary perspective: through its static corollary (the Principle of Minimum Potential Energy), it characterises equilibrium states as potential minima (detailed in Appendix A). Together, these principles motivate the hypothesis that real images, as products of natural dissipative processes, tend to reside near the 'relaxed' basins of the manifold. In contrast, deepfakes are produced by generative models that optimize for statistical similarity to real images but do not explicitly enforce local structural constraints. As a result, synthesized samples may retain microscopic structural inconsistencies, which behave like 'compressed springs' or particles trapped on steep slopes, 'suspended' in high-potential states.

Based on this physical intuition, we introduce the following hypothesis on the topological states of real *vs.* fake data:

Assumption III.2 (**Manifold Stability Hypothesis**)**.** We hypothesize that natural images $\mathbf{x}_{real}$ are typically located near local minima (stable equilibria) of the learned potential surface V. Concretely, for a real latent representation $\mathbf{q}_{real}$, we model $\mathbf{q}_{real}$ as admitting a local neighborhood $\mathcal{N}$ such that:

$$V(\mathbf{q}_{real}) \leq V(\mathbf{q}), \quad \forall \mathbf{q} \in \mathcal{N}. \tag{1}$$

The gradient at equilibrium approximately vanishes: $\|\nabla V(\mathbf{q}_{real})\| \approx 0$. In contrast, we expect deepfake samples $\mathbf{x}_{fake}$ to more frequently occupy high-gradient regions (steep slopes) of V, where $\|\nabla V(\mathbf{q}_{fake})\| \gg 0$[2].

Scope of the Hypothesis. Assumption III.2 is a *distributional* modeling hypothesis about the populations of real and generated images under V, not a per-sample statement about any individual image. The asymmetry we exploit is a statistical tendency: current generative models are trained with distribution-level objectives that do not explicitly enforce local physical regularities, so their outputs are more likely to violate regularity and accumulate elevation under V. We thus treat V as a *probe* for inconsistency with these regularities.

Hamiltonian Dynamics as a Stability Probe. To distinguish the states defined in Assumption III.2, we construct a dynamical system to probe the manifold's geometry. We treat the latent feature $\mathbf{q}$ as a particle system. To capture the varying sensitivity of different semantic features, we introduce a learnable state-conditioned diagonal preconditioner $\mathbf{M}(\mathbf{q})$.

[2]We provide a theoretical motivation linking dissipative mechanics and the Principle of Least Action to this stability hypothesis in Appendix A.

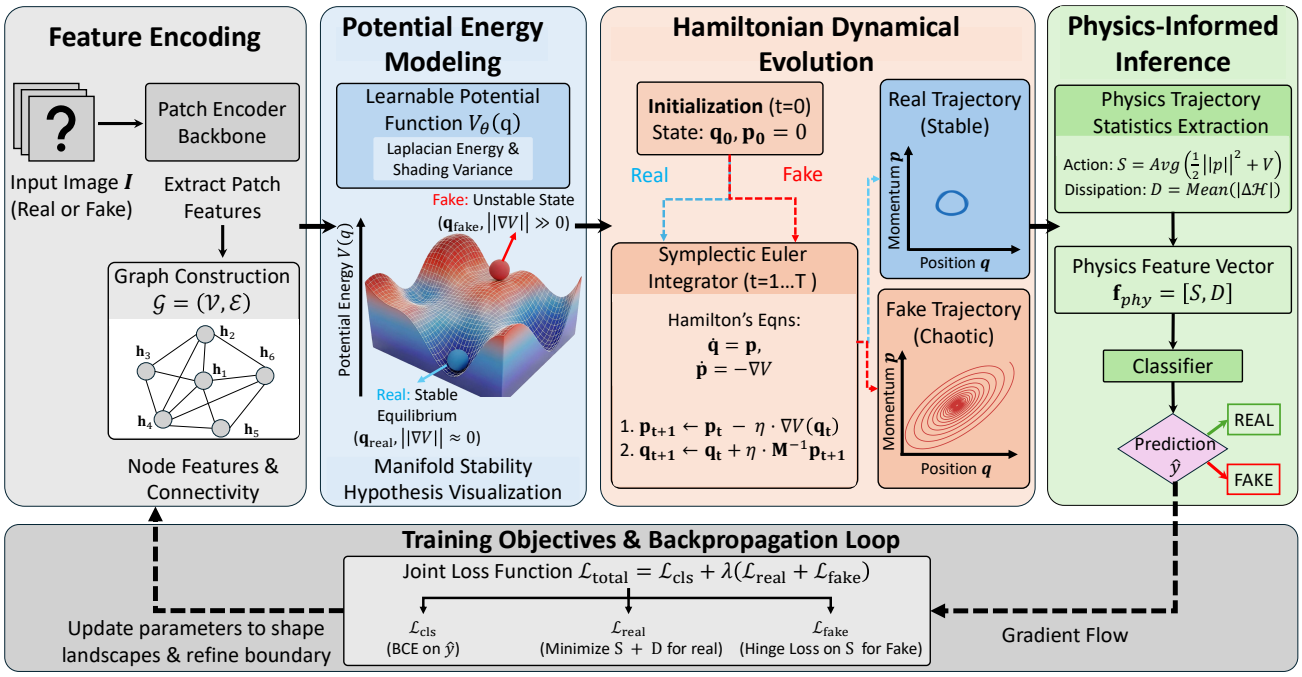

Fig. 3. Overview of the proposed HAAD. We cast deepfake detection as *dynamical stability analysis under a Hamiltonian framework*: **i)** patch features are organized into a spatial graph, then fed into **ii)** a learnable potential V whose landscape encodes physical priors under the Manifold Stability Hypothesis, real samples (blue) are expected to fall near stable equilibria of V while fakes (red) tend to occupy high-gradient unstable states. **iii)** A short-horizon symplectic integrator amplifies these differences into **iv)** trajectory statistics S and D for classification, jointly trained with $\mathcal{L}_{cls}$ and $\mathcal{L}_{phy}$ (bottom loop).

In addition, we introduce a canonical momentum variable $\mathbf{p}$, initialized at zero ($\mathbf{p}_0 = \mathbf{0}$).

Definition III.3 (Hamiltonian System). The total energy of the system, denoted as the Hamiltonian $\mathcal{H}$, is the sum of kinetic energy $T(\mathbf{p})$ and potential energy $V(\mathbf{q})$:

$$\mathcal{H}(\mathbf{q}, \mathbf{p}) = T(\mathbf{p}) + V(\mathbf{q}) = \frac{1}{2}\|\mathbf{p}\|^2 + V(\mathbf{q}). \tag{2}$$

The temporal evolution of the feature state is governed by Hamilton's equations of motion:

$$\frac{d\mathbf{q}}{dt} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}} = \mathbf{p}, \quad \frac{d\mathbf{p}}{dt} = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}} = -\nabla_{\mathbf{q}} V(\mathbf{q}). \tag{3}$$

Physical Interpretation. Equation (3) implies that the **potential gradient** $-\nabla V$, which points toward states more consistent with the physics-motivated regularities, acts as a physical force. Since we initialize at rest ($\mathbf{p}_0 = \mathbf{0}$), any movement is purely driven by this internal tension. Under Assumption III.2, real images (where $\nabla V \approx 0$) are expected to remain approximately stationary, while deepfakes (where $\|\nabla V\| \gg 0$) are expected to experience immediate acceleration, converting potential energy into kinetic momentum.

Proposition III.4 (Divergence of Hamiltonian Action). *Consider the evolution of a particle over a small time interval $t \in [0, \tau]$ starting from rest ($\mathbf{p}_0 = \mathbf{0}$). Under Assumption III.2, the kinetic energy T for a deepfake sample exhibits quadratic growth over time, whereas for a real sample, it remains negligible. As a result, the accumulated Action S satisfies:*

$$S(\mathbf{q}_{fake}) \gg S(\mathbf{q}_{real}). \tag{4}$$

Proof Sketch. We analyze the short-term dynamics via a first-order Taylor expansion of the potential around $\mathbf{q}_0$: $V(\mathbf{q}) \approx V(\mathbf{q}_0) + \nabla V(\mathbf{q}_0)^\top(\mathbf{q} - \mathbf{q}_0)$. Under Assumption III.2, a real image is expected to initialize near a local minimum where $\|\nabla V(\mathbf{q}_0)\| \approx 0$. Hamilton's equations imply momentum rate $\dot{\mathbf{p}} = -\nabla V \approx \mathbf{0}$. Therefore, the particle remains approximately stationary ($\mathbf{p}(t) \approx \mathbf{0}$), yielding minimal Action.

In contrast, a deepfake sample is expected to more often initialize on a slope with significant gradient $\|\nabla V(\mathbf{q}_0)\| = C > 0$. Within the small interval τ, we approximate the gradient force as constant. Integrating $\dot{\mathbf{p}} \approx -C\mathbf{u}$ yields the momentum $\|\mathbf{p}(t)\| \approx Ct$. Under the canonical kinetic energy $T(\mathbf{p}) = \frac{1}{2}\|\mathbf{p}\|^2$ of Definition 2, this grows quadratically:

$$T(t) = \tfrac{1}{2}\|\mathbf{p}(t)\|^2 \approx \tfrac{1}{2}C^2t^2 = \mathcal{O}(t^2). \tag{5}$$

Since S accumulates the total energy ($\mathcal{H} = T + V$) over time, this quadratic surge in T leads to a larger Action score for fake samples than for real samples under Assumption III.2[3]. □

This proposition implies that by simulating Hamiltonian evolution, we can **amplify** the subtle static differences (gradients) into macroscopic dynamic discrepancies (kinetic energy/action), which we use directly as detection cues.

Remark III.5 (**Testable Empirical Prediction**). Proposition III.4 turns the Manifold Stability Hypothesis from a modeling choice into an operational prediction: if the hypothesis describes the real and fake populations well, the trajectory statistic S should be consistently smaller for real samples than for fake samples, not only on the training distribution but also on unseen datasets, unseen manipulations, and unseen generators. This makes the hypothesis indirectly *falsifiable*: if

[3]We provide the detailed derivation in Appendix B.

transfer to unseen distributions failed, the distributional form of the hypothesis would lose empirical support. We provide the cross-dataset, cross-manipulation, and cross-generator experiments in Section IV as the empirical test of this prediction.

B. The HAAD Framework

As illustrated in Figure 3, our HAAD framework contains three key components to translate the theoretical formulation into a differentiable architecture: i) a learnable potential function V, ii) a numerical solver to simulate the continuous dynamics, and iii) physics-informed metrics for detection.

1) Potential Energy Modeling: In statistical physics, potential energy quantifies the tension within a system derived from particle interactions. Extending this to the visual domain, we model an image as a system of *interacting patches*. To instantiate the V introduced in Section III-A, we construct a learnable potential energy surface $V(\mathbf{q};\theta)$, decomposed into two physics-inspired components: Geometric Potential (penalizing structural discontinuities) and Photometric Potential (penalizing illumination inconsistencies).

Physical State Projection. Specifically, given the patch features $\mathbf{x} \in \mathbb{R}^{N\times D_{in}}$ from the backbone, we project them into a unified physical state space. In particular, we employ lightweight heads to estimate the latent position state $\mathbf{q} \in \mathbb{R}^{N\times D}$, alongside three intrinsic physical properties: surface normal $\mathbf{n} \in \mathbb{R}^{N\times 3}$, albedo $\boldsymbol{\rho} \in \mathbb{R}^{N\times 1}$, and global lighting $\mathbf{l} \in \mathbb{R}^3$. The two groups serve distinct roles in the potential decomposition: $\mathbf{q}$ provides the spatial patch representation consumed by the *geometric* component (Dirichlet energy over the patch graph), while $(\mathbf{n}, \boldsymbol{\rho}, \mathbf{l})$ provide the illumination decomposition bases consumed by the *photometric* component. All four projection heads are lightweight linear layers trained *end-to-end* through the total loss $\mathcal{L}_{total}$ (detailed in Section III-C), with no separate auxiliary supervision. As a result, $(\mathbf{n}, \boldsymbol{\rho}, \mathbf{l})$ should be understood as *learned decomposition bases* that make the photometric component of V sensitive to illumination inconsistencies across patches, rather than as calibrated surface properties. This end-to-end design allows the photometric bases to optimize jointly with the detection objective, without requiring separate physical supervision.

Geometric Potential (V_{geo}). To capture structural smoothness, we define a spatial graph $\mathcal{G}$ over the patch grid and compute the graph Laplacian $\mathbf{L} \in \mathbb{R}^{N\times N}$. The geometric potential is formulated as the *Dirichlet energy* of the signal:

$$\begin{aligned} V_{geo}(\mathbf{q}) &= \frac{1}{N}\mathrm{tr}(\mathbf{q}^\top \mathbf{L}\mathbf{q}) \\ &= \frac{1}{N}\sum_{(i,j)\in\mathcal{E}} w_{ij}\|\mathbf{q}_i - \mathbf{q}_j\|^2. \end{aligned} \tag{6}$$

Physically, this term mimics a 'spring force' between neighboring patches. Under this model, real images tend to exhibit smooth feature transitions (low Dirichlet energy), whereas deepfakes often contain high-frequency splicing artifacts or boundary discontinuities, triggering a high energy penalty.

Photometric Potential (V_{photo}). To introduce a heuristic photometric coherence prior, we use a Lambertian-inspired shading constraint, where the potential is defined as the variance of the reconstructed shading field:

$$V_{photo}(\mathbf{n}, \boldsymbol{\rho}, \mathbf{l}) = \mathrm{Var}\left(\boldsymbol{\rho} \odot \mathrm{ReLU}(\mathbf{n}\cdot\mathbf{l})\right). \tag{7}$$

This term acts as a heuristic global coherence regularizer by penalizing high-frequency shading jitter under a shared lighting decomposition, thereby encouraging lighting patterns that are more self-consistent across patches.

Total Potential Surface. The final potential energy is a nonnegative weighted combination of these constraints:

$$V(\mathbf{q}) = \lambda_{geo}V_{geo}(\mathbf{q}) + \lambda_{photo}V_{photo}(\mathbf{n}, \boldsymbol{\rho}, \mathbf{l}), \tag{8}$$

where $\lambda_{geo}, \lambda_{photo} \geq 0$ are scalar coefficients balancing the two regularizers. Since both V_{geo} and V_{photo} are nonnegative by construction, we interpret V up to an additive baseline and rely on relative energy differences rather than an absolute zero level in the theory below.

On the Choice of Physics-inspired Components. The potential $V(\mathbf{q}) = \sum_{k=1}^{K} \lambda_k\, r_k(\mathbf{q})$ can be instantiated from any finite family of differentiable regularity terms $\{r_k\}$. In this work, we adopt $K = 2$ canonical terms: *geometric smoothness* and *photometric consistency*. This selection has three advantages: **(i)** *Lightweight*: they are evaluated directly on patch-level latent features without auxiliary networks; **(ii)** *Empirically violated*: current generative training objectives optimize distribution-level losses rather than such local regularity; and **(iii)** *Complementary*: smoothness acts on spatial gradients of $\mathbf{q}$ while photometric consistency acts on colour-channel relations, so the two terms respond to disjoint failure modes. Additional terms (*e.g.*, frequency-domain priors, colour constancy, or temporal coherence for video) can be added without changes to the Hamiltonian integrator, so K is a design knob rather than a fixed assumption. Table V in Section IV-C isolates the contribution of the two components.

2) Hamiltonian Dynamical Evolution: To probe the stability of the latent state $\mathbf{q}$, we simulate its trajectory by numerically solving Hamilton's equations (Equation (3)). We avoid standard solvers (*e.g.*, Euler method) as they introduce numerical dissipation, causing artificial energy drift that may obscure stability-related signals. Instead, we employ a **Symplectic Integrator** to explicitly preserve the phase-space volume and approximate energy conservation. To account for varying feature sensitivities in the implementation, we introduce a learnable diagonal matrix $\mathbf{M}(\mathbf{q})$ and use $\mathbf{M}(\mathbf{q})^{-1}$ as a *state-conditioned diagonal preconditioner* applied only in the position update of the canonical Hamiltonian system. The momentum update is driven solely by the potential gradient, while the position update rescales the momentum by the local inertia. Using the **Symplectic Euler** scheme with step size η, the state update from t to $t+1$ is:

$$\mathbf{p}_{t+1} = \mathbf{p}_t - \eta\,\nabla_{\mathbf{q}} V(\mathbf{q}_t), \tag{9}$$

$$\mathbf{q}_{t+1} = \mathbf{q}_t + \eta\,\mathbf{M}(\mathbf{q}_t)^{-1}\,\mathbf{p}_{t+1}. \tag{10}$$

Because $\mathbf{M}$ depends on $\mathbf{q}$, a fully general variable-mass Hamiltonian flow would include additional $\mathbf{q}$-derivative terms of $\mathbf{M}(\mathbf{q})^{-1}$ in the momentum update. Our implementation omits these terms for simplicity. In our experiments, across T=4 Symplectic Euler steps and N=1,920

per-step measurements on testing datasets, the median ratio $\|\frac{1}{2}(\partial \mathbf{M}^{-1}/\partial \mathbf{q})\,\mathbf{p}^2\| \,/\, \|\nabla V\|$ is $\sim 10^{-11}$ (max 1.85×10^{-5}), indicating that the omitted variable-mass terms are negligible by multiple orders of magnitude[4]. We therefore describe Equations (9) and (10) as a *Hamiltonian-inspired, symplectic-style stability probe* rather than an exact simulation of a variable-mass Hamiltonian flow. Specifically, the semi-implicit ordering (momentum before position) is strictly symplectic in the ideal constant-mass case (detailed in Appendix C) and is used here as a volume-conserving update family under the same ordering. Additional ablations appear in Section IV-C.

The Gradient as Virtual Perturbation. Since the system initializes at rest ($\mathbf{p}_0 = \mathbf{0}$), the gradient term $-\eta \nabla_{\mathbf{q}} V$ in Equation (9) acts as a **virtual perturbation force**. Under Assumption III.2, for real images residing in flat basins ($\|\nabla V\| \approx 0$), this force is negligible, and the learned preconditioner can further suppress movement, maintaining stationarity. In contrast, for samples on steep slopes (as hypothesized for deepfakes), this significant force works in tandem with the preconditioned position update to deliver an initial 'kick,' driving the state into a divergent trajectory and effectively amplifying the detection signal.

3) Physics-Informed Inference: After evolving the system for T time steps, we extract two trajectory statistics to serve as the detection cues. All metrics are normalized by the number of patches N to ensure resolution invariance.

Hamiltonian Action Score (S). We define the Hamiltonian Action score as the **accumulated dynamical cost** of the system's evolution. It accumulates the total energy magnitude ($\mathcal{H} = T + V$) along the trajectory[5]:

$$S = \frac{1}{T \cdot N} \sum_{t=1}^{T} \mathcal{H}(\mathbf{q}_t, \mathbf{p}_t). \tag{11}$$

Under Assumption III.2, a high S indicates that the sample accumulates substantial energy along its short-horizon trajectory, which is driven by both high potential elevation at the start and growing kinetic energy because of the gradient force.

Energy Dissipation (D). We define D as the mean absolute change in Hamiltonian value along the discrete trajectory:

$$D = \frac{1}{(T-1) \cdot N} \sum_{t=1}^{T-1} |\mathcal{H}_{t+1} - \mathcal{H}_t|. \tag{12}$$

Rather than treating D as a pure physical dissipation term, we use it as an empirical discriminative statistic under the chosen symplectic solver: real trajectories tend to remain approximately energy-conserving under our integrator, while fake trajectories driven by steeper potential gradients tend to produce larger drift. Finally, we concatenate these trajectory features to form the feature vector $\mathbf{f}_{phy} = [S, D]$, which is fed into a linear classifier for the binary decision.

[4]At t=0 the ratio is identically zero since $\mathbf{p}_0$=$\mathbf{0}$. At $t \geq 1$ the median rises to $\sim 10^{-8}$ but remains well below any practical threshold.

[5]We use the term *Hamiltonian Action* (equivalently: *Hamiltonian energy integral* or *accumulated total energy*) to denote $S = \frac{1}{TN}\sum_{t=1}^{T} \mathcal{H}_t = \frac{1}{TN}\sum_t (T_t + V_t)$. This differs from the classical action integral $\int (T - V)\,\mathrm{d}t$, which is minimized by physical paths (Principle of Least Action). We accumulate $T+V$ rather than $T-V$ because the total energy magnitude grows monotonically with any departure from equilibrium, whereas $T-V$ can partially cancel and yields a weaker separator. More details are in Appendix B.

C. Training Objective

Our training objective is a weighted sum of two terms to encourage: i) **Discriminability**, *i.e.*, high detection accuracy, and ii) **Physical Consistency**, *i.e.*, shaping the learned potential landscape to approximate the Manifold Stability Hypothesis,

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{phy}, \tag{13}$$

where λ controls the strength of the physical constraint.

Classification Loss. We pass $\mathbf{f}_{phy}$ through a linear classifier $\sigma(\cdot)$ to compute the logits. Let $\hat{y} = \sigma(\mathbf{f}_{phy})$ be the predicted probability of fake ($y = 1$) *vs.* real ($y = 0$). We minimize the Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{cls} = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]. \tag{14}$$

Physical Regularization Loss. To encourage the Hamiltonian behavior predicted by the Manifold Stability Hypothesis, we add a soft regularizer that shapes the trajectory statistics of real and fake samples in opposite directions:

Stability-Encouragement Term for Real Data: We encourage real images to localize in low-V basins with quasi-conservative dynamics by minimizing their S and D. For the real batch $\mathcal{B}_{real}$:

$$\mathcal{L}_{real} = \mathbb{E}_{\mathbf{x} \in \mathcal{B}_{real}}[S(\mathbf{x}) + D(\mathbf{x})]. \tag{15}$$

Instability-Penalization Term for Fake Data: We penalize fake samples whose trajectories do not diverge sufficiently by applying a hinge loss with margin γ to their Action, effectively encouraging their embeddings toward higher-gradient regions of the potential surface:

$$\mathcal{L}_{fake} = \mathbb{E}_{\mathbf{x} \in \mathcal{B}_{fake}}[\max(0, \gamma - S(\mathbf{x}))]. \tag{16}$$

The total physical regularization is $\mathcal{L}_{phy} = \mathcal{L}_{real} + \mathcal{L}_{fake}$.

IV. Experiment

A. Deepfake Detection

We first performed detection on face manipulation, *i.e.*, standard deepfake benchmarks. Following the previous protocols [19], we trained our model on the **FaceForensics++ (FF++)** [51] (c23) dataset and evaluated on unseen benchmarks. We utilized the Area Under the Curve (AUC) as the metric. For video benchmarks, we uniformly sample 8 frames per video. For video benchmarks, the reported metric is video-level AUC, obtained by averaging per-frame prediction scores within each video, following the unified evaluation protocol [19]. Baseline numbers are sourced from the same protocol where possible, or from the original papers under the identical setting. Our framework employs CLIP ViT-L/14 [33] as the backbone following recent studies [34]. Training is conducted on a single NVIDIA H100 GPU with Adam optimizer (lr $= 2e^{-4}$). The batch size is set to 32. To assess generalization, we evaluated HAAD on two settings: i) **Cross-Dataset:** We employed Celeb-DF++ [13], CDF-v2 [16], DFDC [17], DFD [18], DFo [52], WildDeepfake [53], and FFIW [54] as the testing sets. ii) **Cross-Manipulation:** We applied DF40 [55], which contains unseen forgery types generated within the FF++ domain.

TABLE I

PERFORMANCE COMPARISON WITH SOTA METHODS ON CROSS-DATASET AND CROSS-METHOD EVALUATION. ALL MODELS ARE TRAINED ON FF++ AND TESTED ON DIFFERENT DATASETS USING THE AUC METRIC. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. †: WE RE-IMPLEMENTED THIS MODEL.

Methods	Cross-dataset Evaluation								Cross-method Evaluation								
	CDF++	CDF-v2	DFD	DFDC	DFo	WDF	FFIW	Avg.	UniFace	BleFace	MobSwap	e4s	FaceDan	FSGAN	InSwap	SimSwap	Avg.
F3Net [43]	0.738	0.789	0.844	0.718	0.730	0.728	0.649	0.743	0.809	0.808	0.867	0.494	0.717	0.845	0.757	0.674	0.746
SPSL [44]	0.744	0.799	0.871	0.724	0.723	0.702	0.794	0.769	0.747	0.748	0.885	0.514	0.666	0.812	0.643	0.665	0.710
SRM [45]	0.794	0.840	0.885	0.695	0.722	0.702	0.794	0.767	0.749	0.704	0.779	0.704	0.659	0.772	0.793	0.694	0.732
CORE [46]	0.749	0.809	0.882	0.721	0.765	0.724	0.710	0.762	0.871	0.843	0.959	0.679	0.774	0.958	0.855	0.724	0.833
RECCE [28]	0.808	0.823	0.891	0.696	0.784	0.756	0.711	0.779	0.898	0.832	0.925	0.683	0.848	0.949	0.848	0.768	0.844
SBI [9]	0.734	0.886	0.827	0.717	0.899	0.703	0.866	0.821	0.724	0.891	0.952	0.750	0.594	0.803	0.712	0.701	0.766
UCF [47]	0.761	0.837	0.867	0.742	0.808	0.774	0.697	0.785	0.831	0.827	0.950	0.731	0.862	0.937	0.809	0.647	0.824
IID [48]	0.756	0.838	0.939	0.700	0.808	0.666	0.762	0.789	0.839	0.789	0.888	0.766	0.844	0.927	0.789	0.644	0.811
LSDA [49]	0.727	0.875	0.881	0.701	0.768	0.797	0.724	0.794	0.872	0.875	0.930	0.694	0.721	0.939	0.855	0.793	0.835
Effort† [34]	0.831	0.945	0.951	0.840	0.965	0.827	0.901	0.894	0.950	0.868	0.950	0.987	0.935	0.951	0.915	0.908	0.933
VLFFD† [32]	0.818	0.938	0.940	0.809	0.946	0.831	0.898	0.883	0.875	0.877	0.911	0.921	0.895	0.851	0.837	0.840	0.876
DAID† [6]	0.779	0.909	0.925	0.768	0.884	0.801	0.883	0.850	0.902	0.849	0.874	0.838	0.917	0.935	0.801	0.744	0.858
$\mathcal{X}^2$-DFD† [50]	0.812	0.951	0.959	0.841	0.959	0.841	0.915	0.896	0.874	0.873	0.923	0.912	0.879	0.899	0.820	0.863	0.880
FIA-USA† [7]	0.812	0.935	0.926	0.732	0.874	0.810	0.878	0.852	0.918	0.854	0.842	0.875	0.881	0.863	0.874	0.910	0.877
HAAD (Ours)	**0.878**	**0.963**	**0.975**	**0.868**	**0.979**	**0.851**	**0.932**	**0.921**	**0.971**	**0.874**	**0.964**	**0.990**	**0.939**	**0.968**	**0.930**	**0.917**	**0.944**

TABLE II

PERFORMANCE COMPARISON ON THE GENIMAGE DATASET. ALL MODELS ARE TRAINED ON SDV1.4 AND TESTED ON DIFFERENT SUBSETS USING THE ACC METRIC. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. †: WE RE-IMPLEMENTED THIS MODEL.

Methods	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
CNNSpot [56]	0.528	0.963	0.959	0.501	0.398	0.786	0.534	0.468	0.642
F3Net [43]	0.501	0.997	0.995	0.499	0.500	0.963	0.499	0.499	0.680
UnivFD [31]	0.915	0.964	0.961	0.581	0.734	0.945	0.678	0.577	0.795
NPR [10]	0.810	0.982	0.979	0.769	0.898	0.969	0.841	0.842	0.886
FreqNet [57]	0.896	0.988	0.986	0.668	0.865	0.973	0.758	0.814	0.868
DRCT [58]	0.874	0.893	0.902	0.779	0.891	0.947	0.903	0.859	0.881
Effort† [34]	0.802	0.989	0.983	0.759	0.912	0.968	0.900	0.835	0.894
DDA† [59]	0.902	0.985	0.972	0.809	0.876	0.954	0.735	0.842	0.884
HAAD (Ours)	**0.903**	**0.998**	**0.997**	**0.817**	**0.951**	**0.986**	**0.928**	**0.870**	**0.931**

Experimental design rationale. The cross-domain transfer serves as the empirically independent test of the Manifold Stability Hypothesis: a learned V that captured only dataset-specific artifacts would fail to generalize to unseen generators, whereas physics-motivated regularities that reflect genuine structural properties of image formation should transfer.

Main Results. Table I presents a comparison of HAAD against 14 SoTA detectors. Our method achieves the highest average AUC among the evaluated baselines across the tested settings. Specifically, in the cross-dataset setting, HAAD achieves an average AUC of 0.921, outperforming all of the competing baselines. Furthermore, in the challenging CDF++ benchmark, it surpasses the second-best baseline by approximately 5 pp. In the cross-manipulation setting, HAAD achieves the highest average AUC among the evaluated methods (0.944). These results indicate that the Hamiltonian stability constraints capture forgery-related signals that generalize beyond specific manipulation patterns.

B. Synthetic Image Detection

We utilized the GenImage [21] benchmark. Following standard protocols [34], models are trained on the SDv1.4 subset and evaluated on the remaining unseen subsets. We reported Accuracy (Acc) against 8 baseline models in Table II. HAAD achieves the highest average Acc among the evaluated baselines (0.931). For instance, on the GLIDE subset, our approach surpasses the second-best baseline by approximately 4 percentage points. Furthermore, HAAD attains the highest Acc in the intra-domain setting (*i.e.*, when evaluated in the SDv1.4 source domain). These results indicate that the physics-inspired stability prior generalizes well to unseen generative models in standard transfer settings, despite being trained on a single source domain.

TABLE III

ABLATION STUDIES ON THE CONTRIBUTION OF PHYSICAL COMPONENTS. WE REPORTED THE AUC SCORES ON THE CROSS-DATASET PROTOCOL.

Variant	**Potential V**	**Action (S)**	**Dissipation (D)**	**CDF++**	**DFDC**
Backbone only	-	-	-	0.734	0.732
+Static V	✓	-	-	0.845	0.851
+Static $[V_0, \|\nabla V_0\|]$	✓	-	-	0.848	0.853
+Action S (T=4)	✓	✓	-	0.860	0.858
+Action S (λ=0)	✓	✓	-	0.852	0.854
HAAD (Ours)	✓	✓	✓	**0.878**	**0.868**

TABLE IV

COMPARISON OF DIFFERENT POTENTIAL SURFACE CONSTRUCTION. WE REPORTED THE AUC SCORES ON THE CROSS-DATASET PROTOCOL.

Topology	**CDF++**	**CDF-v2**	**DFD**	**DFDC**	**DFo**	**Avg**
Independent (MLP)	0.855	0.951	0.962	0.846	0.970	0.917
Global (Self-Attn)	0.869	0.958	0.971	0.851	0.977	0.925
Graph (Ours)	**0.878**	**0.963**	**0.975**	**0.868**	**0.979**	**0.933**

C. Ablation Studies

Component Effectiveness Analysis. We ablate five variants in Table III starting from the standalone CLIP ViT-L/14 backbone: **i)** *Static* V: potential as soft regularizer, no trajectory features; **ii)** *Static* $[V_0, \|\nabla V_0\|]$: static potential and gradient magnitude as classifier inputs, no rollout; **iii)** *Action* S: T=4 symplectic rollout with $\mathcal{L}_{phy}$; **iv)** *Action* S, λ=0: same rollout,

TABLE V
ABLATION STUDIES ON THE FACTORIZATION OF THE POTENTIAL SURFACE. WE REPORTED THE AUC SCORES ON THE CROSS-DATASET PROTOCOL. Δ IS THE AVERAGE-AUC DROP (PP).

Variant	λ_{geo}	λ_{photo}	FF++	CDF-v2	DFDC	Δ
smooth-only	✓	–	0.993	0.954	0.844	1.2
photo-only	–	✓	0.994	0.948	0.852	1.1
both (Ours)	✓	✓	**0.995**	**0.963**	**0.868**	0.0

TABLE VI
ABLATION STUDIES ON NUMERICAL SOLVERS. WE REPORTED AUC SCORES ON THE CROSS-DATASET PROTOCOL. IN ADDITION, WE COMPARED THE EFFICIENCY OF DIFFERENT SOLVERS.

Solver	CDF++	CDF-v2	DFD	DFDC	DFo	Time	Mem.
Euler (1st)	0.853	0.940	0.957	0.843	0.944	**12ms**	2480M
RK4 (4th)	0.875	0.959	0.971	0.863	0.975	45ms	3100M
HAAD	**0.878**	**0.963**	**0.975**	**0.868**	**0.979**	14ms	**2485M**

TABLE VII
MASS-MATRIX ABLATION FOR THE POSITION UPDATE. WE REPORTED THE AUC SCORES ON THE CROSS-DATASET PROTOCOL. Δ IS THE AVERAGE-AUC DROP (PP).

Variant	$\mathbf{M}(\mathbf{q})$	CDF-v2	DFDC	FF++	Avg	Δ
Constant mass ($\mathbf{M}=\mathbf{I}$)	–	0.949	0.844	0.992	0.928	1.4
Learned $\mathbf{M}(\mathbf{q})$ (Ours)	✓	**0.963**	**0.868**	**0.995**	**0.942**	0.0

$\mathcal{L}_{cls}$ only; and **v)** *Full HAAD*: complete model. From Table III, we have two main observations: *(1) Every component is indispensable.* Each design choice contributes measurable and consistent gains. For instance, removing only D from Full HAAD degrades CDF++ AUC from 0.878 to 0.860, and replacing $\mathcal{L}_{phy}$ with pure classification supervision (variant iv) likewise causes a consistent drop on both datasets, indicating that the trajectory statistics and physical regularization ($\mathcal{L}_{phy}$) each provide independent discriminative signal. *(2) The rollout is the key amplifier, and D captures complementary trajectory curvature.* Directly classifying with static features (variant ii) adds only a marginal gain over the regularizer-only baseline (variant i), whereas the $T{=}4$ symplectic rollout (variant iii) yields +1.2 pp on CDF++ over the same reference, confirming that trajectory-accumulated kinetic growth is far more discriminative than any single-point gradient readout, consistent with Proposition III.4. Full HAAD (variant v) achieves 0.878 on CDF++ and 0.868 on DFDC, with D contributing the largest single step. A possible reason is that D measures step-to-step Hamiltonian volatility rather than accumulated energy, thereby capturing the curvature of V traversed along the trajectory, which is a complementary to S.

Impact of Potential Energy Architectures. We further justified the choice of constructing the potential surface via **Graph Topology** by comparing it with two alternatives: i) **Independent MLP**, processing patches in isolation, and ii) **Global Self-Attention**, computing fully-connected correlations. Table IV shows that our graph-based potential achieves the highest average AUC among the tested constructions. We attributed this to the physical inductive bias. Unlike global attention or isolated processing, the graph Laplacian explicitly acts as a discrete differential operator. It encourages *local* physical continuity, making it sensitive to high-frequency boundary artifacts and blending discontinuities typical of deepfakes.

Choice of the $K{=}2$ Constraint Decomposition. To validate the two-term decomposition, we ablated the two gates of $V(\mathbf{q}) = \lambda_{geo}V_{geo} + \lambda_{photo}V_{photo}$ while holding every other ingredient fixed. Specifically, three variants are compared: i) **smooth-only** ($\lambda_{geo}{>}0$, $\lambda_{photo}{=}0$), retaining the graph Laplacian but disabling shading; ii) **photo-only** ($\lambda_{geo}{=}0$, $\lambda_{photo}{>}0$), retaining Lambertian residuals but disabling geometric smoothness; and iii) **both** (Ours), which activates both terms. As shown in Table V, disabling the photometric term (smooth-only) drops the average AUC by 1.2 pp, while disabling the geometric term (photo-only) drops it by 1.1 pp. We draw two conclusions from these observations: i) Each term contributes independently, and neither alone carries the full improvement. ii) The geometric and photometric regularities target disjoint failure modes, with each providing discriminative signal the other cannot supply.

Choice of Numerical Integrator. The Symplectic Euler integrator preserves phase-space volume and maintains approximate energy conservation, keeping the trajectory statistic D sensitive to data-dependent potential gradients. To verify this advantage, we compared our **Symplectic Euler** against two standard alternatives: i) the first-order **Euler Method** and ii) the fourth-order **Runge-Kutta (RK4)**. As shown in Table VI, the Euler method yields the poorest performance due to its inherent lack of energy conservation, introducing significant artificial drift. Although RK4 offers higher precision, it is non-symplectic and incurs substantial cost (45 ms, $+650$ MB) due to multiple gradient evaluations. In contrast, the proposed Symplectic Euler achieves the best trade-off. It outperforms RK4 while adding only 2.5 ms and 35 MB over the CLIP ViT-L/14 backbone (11.5 ms / 2450 MB). Moreover, its symplectic property preserves phase-space volume, which reduces solver-induced energy drift relative to non-symplectic methods. While this does not eliminate all numerical effects, it makes dissipation (D) a more reliable empirical discriminative signal.

Mass-Matrix Sensitivity. The state-conditioned diagonal preconditioner $\mathbf{M}(\mathbf{q})^{-1}$ in the position update (Equation (10)) assigns different effective step sizes to different patch feature dimensions, allowing the rollout dynamics to adaptively amplify position updates along the dimensions most sensitive to fake-related instability while damping updates along stable, consistent ones. To verify this contribution, we compare full HAAD against a variant with constant identity mass ($\mathbf{M}{=}\mathbf{I}$), holding all other components fixed. As reported in Table VII, the learned preconditioner consistently improves detection: CDF-v2 AUC increases from 0.949 to 0.963 and the three-set average from 0.928 to 0.942 (+1.4 pp). This indicates that $\mathbf{M}(\mathbf{q})^{-1}$ successfully learns to focus the trajectory dynamics on the most discriminative feature directions, producing sharper trajectory statistics than uniform-step dynamics alone.

Impact of Physical Regularization Weight (λ). The weight λ controls the balance between classification and physical consistency. Table VIII shows that performance peaks at $\lambda = 1.0$ (0.878). Values below this threshold provide insufficient con-

TABLE VIII
SENSITIVITY TO PHYSICAL LOSS WEIGHT λ ON CDF++.

λ	0.0	0.1	0.3	0.5	**1.0**	2.0	3.0	5.0
AUC	0.845	0.862	0.867	0.871	**0.878**	0.874	0.869	0.860

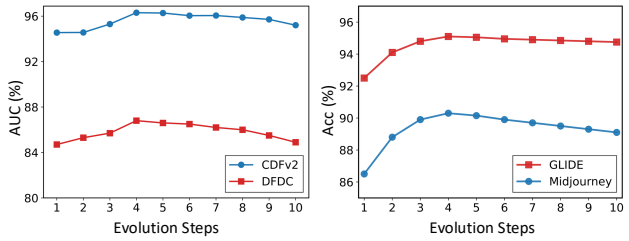

Fig. 4. Ablation studies on evolution steps. The detection performance for varying T is reported on (a) deepfake benchmarks and (b) GenImage subsets.

straints to shape the manifold, while larger values ($\lambda > 1.0$) cause the regularization term to dominate the optimization, leading to suboptimal discrimination.

Sensitivity to Evolution Steps. The number of evolution steps T controls how far the trajectory probe extends. To evaluate sensitivity to trajectory length, we varied T from 1 to 10 and reported the performance in Figure 4. We observed that at $T = 1$, the model behaves similarly to a standard gradient-based regularization, yielding suboptimal performance. The detection accuracy improves rapidly as T increases, peaking around $T = 4$. This indicates that a short-term trajectory is sufficient to reveal the stability gap: real samples remain bounded while fakes begin to diverge. Further increasing T leads to a slight saturation. We attributed this to the local nature of the learned potential, *i.e.*, long trajectories may drift into poorly constrained regions far from the valid distribution.

D. Qualitative Analysis

Dynamic Evolution Trajectories. Figure 5a illustrates the normalized energy trajectories, plotting the median value (solid line) and the interquartile range (shaded region). The results reveal a clear contrast: real images (blue) maintain *bounded, quasi-stable orbits*, consistent with placement in deep potential basins dominated by restorative forces. In contrast, deepfake samples (red) exhibit rapid energy divergence even within limited steps ($T = 4$). This is consistent with the hypothesis that forged content occupies unstable high-potential states, where significant gradient forces act as persistent accelerators, driving the particle state away from equilibrium.

Separability of the Action Statistic. Figure 5b displays the density histogram of the relative Hamiltonian Action ($S - S_{real}$), where the total area under each curve is normalized to 1. Real samples (blue) form a sharp, narrow peak clustered tightly around zero. This is consistent with real images residing in low-energy basins with minimal fluctuation under the learned potential. In contrast, fake images (red) exhibit a *heavy-tailed distribution* spanning a wide range of large positive relative-action values. This dispersion reflects the diverse nature of generative artifacts, which place samples at various unstable elevations on the potential surface. The

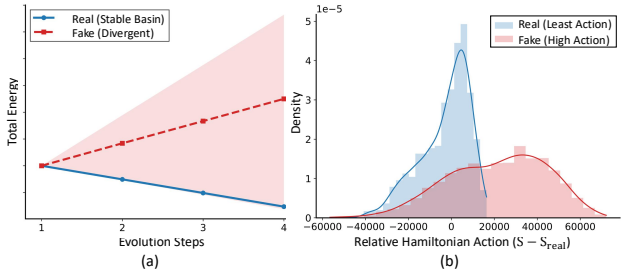

Fig. 5. (a) Energy evolution trajectories. Real images (blue) exhibit bounded stability, whereas deepfakes (red) demonstrate a distinct divergent trend. We plotted the median normalized energy with the interquartile range (shaded). (b) Distribution of relative Hamiltonian Action ($S - S_{real}$). The sharp peak for real data *vs.* the heavy-tailed distribution for fakes highlights the discriminative power of this stability-based metric.

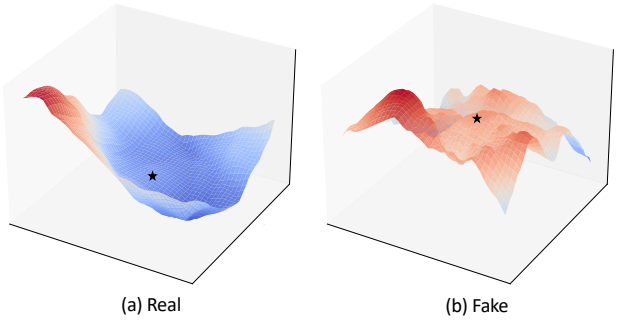

Fig. 6. Potential landscape visualization. We visualize the learned energy surface by projecting the latent neighborhood onto random orthogonal directions. (a) Around a representative real face the learned surface forms a smooth basin. (b) Around a representative fake face the learned surface contains a high-curvature steep region associated with larger trajectory responses.

distinct separation between these two distributions indicates that S serves as an effective linear separator for detection.

3D Potential Landscape Visualization. To visualize the learned energy landscape, we projected the local neighborhood of samples onto random orthogonal directions. The landscape around a real face (Figure 6a) manifests as a *smooth, convex basin*, with the data point located at a stable local minimum. This is consistent with natural images residing near equilibria under the learned potential. In contrast, the landscape for a deepfake (Figure 6b) reveals a *high-curvature steep region* rather than a stable basin. The forged sample is located in an unstable high-potential area with large local gradients.

E. Robustness and Scope Analysis

To delineate HAAD's practical scope, we evaluated two deployment-relevant axes: i) Common post-processing corruptions at test time, and ii) Backbone choice.

Common Post-processing Corruptions. Real-world forensic pipelines re-encode, resize and re-share images before a detector sees them. To simulate this, we evaluated HAAD and three representative baselines with applying a standard set of test-time corruptions: JPEG compression at two quality levels (Q=70 and Q=50), Gaussian blur (σ=3), and additive Gaussian noise (σ^2=0.01). All detectors remain the same clean-trained checkpoint, without any corruption-aware retraining. Table IX shows that HAAD remains the strongest detector

TABLE IX
COMMON-CORRUPTION ROBUSTNESS. WE REPORTED AUC ON CDF-V2. $\Delta_{\max}$ IS THE WORST-CASE AUC DROP RELATIVE TO CLEAN.

Method	Clean	JPEG Compression Q=70	JPEG Compression Q=50	Gaussian Blur (σ=3)	Gaussian Noise (σ^2=0.01)	$\Delta_{\max} \downarrow$
EfficientNet	0.646	0.582	0.522	0.543	0.519	0.127
SBI	0.886	0.846	0.753	0.784	0.762	0.133
Effort	0.945	0.873	0.801	0.826	0.798	0.147
HAAD	**0.963**	**0.925**	**0.904**	**0.882**	**0.857**	**0.106**

TABLE X
BACKBONE SENSITIVITY EVALUATION. WE REPORTED THE AUC SCORES OF TWO DIFFERENT BACKBONES ON CDF-V2.

Backbone	Base	+HAAD	Δ
ResNet-50 [60]	0.683	**0.806**	+0.123
ViT-B/16	0.915	**0.948**	+0.033

among the evaluated models under every tested corruption, and that its worst-case degradation is smaller than that observed for the baselines. This is consistent with the central framing of HAAD as a *stability probe*: if the discriminative signal is the short-horizon stability of the learned potential rather than a fragile high-frequency artifact, then post-processing, which perturbs local pixel statistics more than it perturbs the smoothed manifold geometry, would be expected to degrade HAAD less than artifact-based detectors.

Backbone Sensitivity. To evaluate whether the Hamiltonian stability probe generalizes across architectures, we applied the HAAD head to two additional backbones: ResNet-50 [60] and ViT-B/16. As shown in Table X, HAAD improves every backbone. For instance, it improves the ResNet-50 baseline by 12.3 pp. This indicates that S and D serve as an effective detection prior across diverse backbone architectures.

V. CONCLUSION AND DISCUSSION

In this paper, we propose a new perspective on deepfake detection: moving from static pattern recognition to **dynamical stability analysis**. By conceptualizing the latent features as a physical energy landscape, we propose **HAAD**, a framework that complements artifact-based detection by probing dynamical stability across generative architectures. Our core contribution is the operationalization of the Manifold Stability Hypothesis via a symplectic evolution scheme. Furthermore, the short-horizon symplectic evolution amplifies microscopic structural irregularities into measurable trajectory differences, summarized by the Hamiltonian Action and Energy Dissipation. Experimental results demonstrate that the proposed stability-based prior generalizes across generators and datasets, outperforming several SoTA baselines.

Limitations and Societal Considerations. HAAD is a stability-based probe and its effectiveness depends on the learned potential capturing structural regularities that current generators do not explicitly enforce. The Manifold Stability Hypothesis is distributional and does not provide per-sample guarantees. Like all published detectors, HAAD participates in the ongoing detector-generator arms race: future generators may learn to suppress the measured stability cues, which motivates continued development of physics-grounded detection priors. Because forensic false positives can carry reputational and legal consequences, we view HAAD as a decision-support signal that should be used alongside provenance analysis, contextual evidence, and human review.

REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[2] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," in *ICML*, 2024, pp. 1–13.

[3] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *CVPR*, 2016, pp. 2387–2395.

[4] J. Cheng, Z. Yan, Y. Zhang, Y. Luo, Z. Wang, and C. Li, "Can we leave deepfake data behind in training deepfake detector?" in *NeurIPS*, 2024, pp. 21 979–21 998.

[5] C. Hong, Y. Hsu, and T. Liu, "Contrastive learning for deepfake classification and localization via multi-label ranking," in *CVPR*, 2024, pp. 17 627–17 637.

[6] H. Cheng, M.-H. Liu, Y. Guo, T. Wang, L. Nie, and M. Kankanhalli, "Fair deepfake detectors can generalize," in *NeurIPS*, 2025.

[7] L. Ma, Z. Yan, J. Xu, Y. Chen, Q. Guo, Z. Bi, Y. Liao, and H. Lin, "From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes," in *NeurIPS*, 2025.

[8] R. Xia, D. Liu, J. Li, L. Yuan, N. Wang, and X. Gao, "Mmnet: multi-collaboration and multi-supervision network for sequential deepfake detection," *IEEE TIFS*, vol. 19, pp. 3409–3422, 2024.

[9] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *CVPR*, 2022, pp. 18 699–18 708.

[10] C. Tan, H. Liu, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection," in *CVPR*, 2024, pp. 28 130–28 139.

[11] Y.-H. Han, T.-M. Huang, K.-L. Hua, and J.-C. Chen, "Towards more general video-based deepfake detection through facial component guided adaptation for foundation model," in *CVPR*, 2025, pp. 22 995–23 005.

[12] H. Cheng, Y. Guo, T. Wang, L. Nie, and M. Kankanhalli, "Diffusion facial forgery detection," in *ACM MM*, 2024, p. 5939–5948.

[13] Y. Li, D. Zhu, X. Cui, and S. Lyu, "Celeb-DF++: A Large-Scale challenging video DeepFake benchmark for generalizable forensics," *arXiv preprint arXiv:2507.18015*, 2025.

[14] M. Lutter, C. Ritter, and J. Peters, "Deep Lagrangian Networks: Using physics as model prior for deep learning," *arXiv preprint arXiv:1907.04490*, 2019.

[15] M. Cranmer, S. Greydanus, S. Hoyer, P. Battaglia, D. Spergel, and S. Ho, "Lagrangian neural networks," *arXiv preprint arXiv:2003.04630*, 2020.

[16] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *CVPR*, 2020, pp. 3204–3213.

[17] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[18] Google AI Blog, "Contributing Data to Deepfake Detection Research," 2019. [Online]. Available: https://research.google/blog/contributing-data-to-deepfake-detection-research/

[19] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A comprehensive benchmark of deepfake detection," in *NeurIPS*, 2023, pp. 4534–4565.

[20] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *CVPR*, 2023, pp. 8568–8577.

[21] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, "GenImage: A million-scale benchmark for detecting AI-generated image," in *NeurIPS*, 2023, pp. 77 771–77 782.

[22] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for diffusion-generated image detection," in *ICCV*, 2023, pp. 22 445–22 455.

[23] R. Xia, D. Zhou, D. Liu, L. Yuan, S. Wang, J. Li, N. Wang, and X. Gao, "Advancing generalized deepfake detector with forgery perception guidance," in *ACM MM*, 2024, pp. 6676–6685.

[24] M.-H. Liu, H. Cheng, T. Wang, X. Luo, and X.-S. Xu, "Learning real facial concepts for independent deepfake detection," in *IJCAI*, 2025, pp. 1585–1593.

[25] S. Jia, C. Ma, T. Yao, B. Yin, S. Ding, and X. Yang, "Exploring frequency adversarial attacks for face forgery detection," in *CVPR*, 2022, pp. 4093–4102.

[26] X. Wu, Z. Xie, Y. Gao, and Y. Xiao, "Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features," in *ICASSP*, 2020, pp. 2952–2956.

[27] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *CVPR*, 2020, pp. 5000–5009.

[28] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *CVPR*, 2022, pp. 4103–4112.

[29] H. Wu, J. Zhou, and S. Zhang, "Generalizable synthetic image detection via language-guided contrastive learning," *arXiv preprint arXiv:2305.13800*, 2023.

[30] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao, "Forgery-aware adaptive transformer for generalizable synthetic image detection," in *CVPR*, 2024, pp. 10 770–10 780.

[31] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *CVPR*, 2023, pp. 24 480–24 489.

[32] K. Sun, S. Chen, T. Yao, Z. Zhou, J. Ji, X. Sun, C.-W. Lin, and R. Ji, "Towards general visual-linguistic face forgery detection," in *CVPR*, 2025, pp. 19 576–19 586.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[34] Z. Yan, J. Wang, P. Jin, K.-Y. Zhang, C. Liu, S. Chen, T. Yao, S. Ding, B. Wu, and L. Yuan, "Orthogonal subspace decomposition for generalizable ai-generated image detection," in *ICML*, 2025, pp. 1–13.

[35] X. Liao, Y. Wang, T. Wang, K. P. Chow, and S. Lyu, "FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks," *IEEE TCSVT*, vol. 33, no. 12, pp. 7236–7251, 2023.

[36] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. Snavely, "Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting," in *CVPR*, 2021, pp. 5453–5462.

[37] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image," in *CVPR*, 2020, pp. 2475–2484.

[38] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *NeurIPS*, 2020, pp. 21 464–21 475.

[39] S. Greydanus, M. Dzamba, and J. Yosinski, "Hamiltonian neural networks," in *NeurIPS*, 2019, pp. 15 353–15 363.

[40] D. Duvenaud, J. Wang, J. Jacobsen, K. Swersky, M. Norouzi, and W. Grathwohl, "Your classifier is secretly an energy based model and you should treat it like one," in *ICLR*, 2020.

[41] Y. D. Zhong, B. Dey, and A. Chakraborty, "Symplectic ODE-Net: Learning Hamiltonian dynamics with control," *arXiv preprint arXiv:1909.12077*, 2019.

[42] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities." *PNAS*, vol. 79, no. 8, pp. 2554–2558, 1982.

[43] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *ECCV*, 2020, pp. 86–103.

[44] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *CVPR*, 2021, pp. 772–781.

[45] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *CVPR*, 2021, pp. 16 317–16 326.

[46] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao, "Core: Consistent representation learning for face forgery detection," in *CVPRW*, 2022, pp. 12–21.

[47] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: uncovering common features for generalizable deepfake detection," in *ICCV*, 2023, pp. 22 355–22 366.

[48] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *CVPR*, 2023, pp. 3994–4004.

[49] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, "Transcending forgery specificity with latent space augmentation for generalizable deepfake detection," in *CVPR*, 2024, pp. 8984–8994.

[50] Y. Chen, Z. Yan, G. Cheng, K. Zhao, S. Lyu, and B. Wu, "X2-dfd: A framework for explainable and extendable deepfake detection," in *NeurIPS*, 2025.

[51] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *ICCV*, 2019, pp. 1–11.

[52] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *CVPR*, 2020, pp. 2889–2898.

[53] B. Zi, M. Chang, J. Chen, X. Ma, and Y. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *ACM MM*, 2020, pp. 2382–2390.

[54] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild," in *CVPR*, 2021, pp. 5778–5788.

[55] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu *et al.*, "DF40: Toward next-generation deepfake detection," in *NeurIPS*, 2024, pp. 29 387–29 434.

[56] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *CVPR*, 2020, pp. 8695–8704.

[57] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning," in *AAAI*, 2024, pp. 5052–5060.

[58] B. Chen, J. Zeng, J. Yang, and R. Yang, "Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images," in *ICML*, 2024, pp. 7621–7639.

[59] R. Chen, J. Xi, Z. Yan, K.-Y. Zhang, S. Wu, J. Xie, X. Chen, L. Xu, I. Guan, T. Yao *et al.*, "Dual data alignment makes AI-generated image detector easier generalizable," in *NeurIPS*, 2025.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[61] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2004, no. 14.

APPENDIX A
THEORETICAL MOTIVATION FOR THE MANIFOLD STABILITY HYPOTHESIS

In this section, we provide a theoretical motivation for the Manifold Stability Hypothesis proposed in Section III-A. We connect the Principle of Least Action (PLA) to the static geometric properties of the image manifold, motivating the intuition that real and fake data may occupy distinct topological states (valleys *vs.* peaks). We emphasize that the following derivation is conditional on assumptions about image formation and does not constitute a universal proof.

A. From Least Action to Minimum Potential Energy

The **PLA** states that the physical evolution of a system between two states follows a path that makes the action functional stationary ($\delta\mathcal{S} = 0$). While PLA primarily governs dynamic trajectories, it has a direct corollary for the static states of macroscopic systems formed under natural laws: the **Principle of Minimum Potential Energy**.

Natural images can be viewed as terminal states of complex physical processes (*e.g.*, photon transport, sensor integration, and chemical development). These processes are generally **dissipative**, *i.e.*, they involve energy exchange with the environment and tend toward thermodynamic equilibrium.

Remark (Relaxation to Equilibrium via Langevin Dynamics). Consider the image formation process as a stochastic dynamical system governed by Langevin dynamics:

$$\frac{d\mathbf{q}}{dt} = -\nabla V(\mathbf{q}) - \gamma\mathbf{p} + \sqrt{2\gamma\beta^{-1}}\boldsymbol{\xi}(t), \tag{17}$$

where γ represents the dissipation (friction) coefficient, and $\boldsymbol{\xi}(t)$ is Gaussian noise. Over time, dissipative terms drain the kinetic energy, and the system naturally settles into states where the net force vanishes ($\nabla V \approx 0$) and the potential energy is locally minimized.

Therefore, we hypothesize that **real images** ($\mathbf{x}_{real}$) tend to represent these 'relaxed' states. They are expected to reside near the **basins of attraction** (local minima) of the underlying natural potential surface V, characterized by structural coherence and minimal internal tension.

B. Deepfakes as High-Tension States

In contrast, deepfake generation (*e.g.*, GANs, Diffusion Models) is not a natural dissipative relaxation but a **statistical optimization**. The generator G aims to map a latent code $\mathbf{z}$ to a target distribution p_{data} by minimizing a statistical divergence $\mathcal{D}$ (*e.g.*, Jensen-Shannon or KL divergence).

However, minimizing statistical divergence $\mathcal{D}$ does not imply minimizing the physical potential V.

- **Forced Fitting:** To satisfy the semantic constraints (*e.g.*, 'identity of Person A' + 'expression of Person B'), the generator must fuse disjoint features that may not naturally coexist.
- **Structural Tension:** This fusion creates microscopic inconsistencies, such as unnatural pixel correlations or discontinuity in high-frequency domains. In our Hamiltonian framework, these inconsistencies manifest as **structural tension** (high potential energy).

Physical Analogy. Conceptually, a deepfake can be likened to a **compressed spring**: an external optimization pins the sample to a specific point, while the learned potential's gradient $-\nabla V$ acts as a restorative force pulling against that placement. The system therefore has high stored potential energy. Unlike real images, which rest in basins of the potential well, deepfakes under this view are suspended on **steep slopes** of V, where the gradient magnitude is significant: $\|\nabla V\| \gg 0$.

C. Hamiltonian Dynamics as a Stability Probe

Since the potential function V is high-dimensional and implicit, we cannot directly measure the 'elevation' of a sample. Instead, we use **Hamiltonian Dynamics** as a probe to test the local stability of the state.

We treat the feature vector $\mathbf{q}$ as a particle initialized at rest; the potential gradient then induces virtual momentum during the rollout. The subsequent evolution follows the Hamiltonian $\mathcal{H} = T + V$:

$$\frac{d\mathbf{p}}{dt} = -\nabla V(\mathbf{q}). \tag{18}$$

This leads to two distinct dynamic behaviors:

1) **Oscillation (Real):** For a particle at a local minimum ($\nabla V \approx 0$), the force acts as a restorative force, keeping the particle confined within the basin. The kinetic energy T remains bounded and small. Consequently, the accumulated **Hamiltonian Action (Cost)** remains close to the real-sample baseline (equivalently, $S - S_{real} \approx 0$ in the local comparison).
2) **Divergence (Fake):** For a particle on a slope ($\|\nabla V\| \gg 0$), the non-zero gradient acts as a persistent accelerating force. The stored potential energy is rapidly converted into kinetic energy ($T \propto t^2$), causing the particle to accelerate away from the initial state. This results in a larger accumulated **Hamiltonian Action**.

By simulating this short-term evolution, our HAAD framework effectively amplifies the microscopic static tension into a macroscopic dynamic signal, providing a complementary discriminator that is less dependent on specific visual artifacts than static detection methods.

APPENDIX B
PROOFS AND DERIVATIONS FOR PROPOSITION III.4

In this section, we provide the detailed derivation for Proposition III.4 (Divergence of Hamiltonian Action) presented in the main text. We analyze the short-term evolution of the latent state $\mathbf{q}$ under the proposed Hamiltonian system and bound the growth of the Hamiltonian Action S for real and fake samples, respectively. This derivation is conditional on the Manifold Stability Hypothesis (Assumption 3.1); it formalizes the consequence of the hypothesis but does not prove that the hypothesis holds universally.

A. Preliminaries

Recall the canonical Hamiltonian of our system (Definition III.3 in the main text):

$$\mathcal{H}(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\|\mathbf{p}\|^2 + V(\mathbf{q}). \tag{19}$$

The dynamics follow Hamilton's equations:

$$\dot{\mathbf{q}} = \mathbf{p}, \quad \dot{\mathbf{p}} = -\nabla V(\mathbf{q}). \tag{20}$$

In the implementation (Section III-B2), the position update is additionally preconditioned by a state-conditioned diagonal matrix $\mathbf{M}(\mathbf{q})^{-1}$ that rescales per-feature step sizes; this preconditioner modulates the constant in the kinetic-growth bound below but does not affect the $\mathcal{O}(\tau^2)$ separation between real and fake samples. We therefore analyze the canonical system here for clarity. We consider a short time interval $t \in [0, \tau]$, initialized at rest: $\mathbf{q}(0) = \mathbf{q}_0$ and $\mathbf{p}(0) = \mathbf{0}$.

B. Proof of Divergence

To analyze the local behavior, we perform a second-order Taylor expansion of the potential energy $V(\mathbf{q})$ around the initial state $\mathbf{q}_0$. For any state $\mathbf{q}$ in the local neighborhood:

$$\begin{aligned} V(\mathbf{q}) \approx V(\mathbf{q}_0) + \nabla V(\mathbf{q}_0)^\top (\mathbf{q} - \mathbf{q}_0) \\ + \frac{1}{2}(\mathbf{q} - \mathbf{q}_0)^\top \mathbf{H}(\mathbf{q}_0)(\mathbf{q} - \mathbf{q}_0), \end{aligned} \tag{21}$$

where $\mathbf{H}(\mathbf{q}_0)$ is the Hessian matrix of V at $\mathbf{q}_0$.

1) Case 1: Real Samples (Stable Equilibrium): Under the **Manifold Stability Hypothesis** (Assumption III.2), a real sample $\mathbf{q}_{real}$ is hypothesized to reside near a local minimum. This neighborhood implies two conditions:

1) **Vanishing Gradient:** $\|\nabla V(\mathbf{q}_{real})\| \approx 0$.
2) **Positive Definite Hessian:** The Hessian $\mathbf{H}(\mathbf{q}_{real})$ is positive semi-definite (convex basin).

Substituting $\nabla V \approx \mathbf{0}$ into Hamilton's equations:

$$\dot{\mathbf{p}}(0) = -\nabla V(\mathbf{q}_{real}) \approx \mathbf{0}. \tag{22}$$

Since the initial momentum $\mathbf{p}(0) = \mathbf{0}$ and the initial force is zero, the system remains in equilibrium at first-order approximation. The kinetic energy $T(t)$ remains negligible:

$$T(t) = \frac{1}{2}\|\mathbf{p}(t)\|^2 \approx 0. \tag{23}$$

Consequently, the accumulated Hamiltonian Action S is dominated by the local baseline potential of the real sample:

$$S_{real} \approx \frac{1}{\tau}\int_0^\tau V(\mathbf{q}_{real})dt \approx V(\mathbf{q}_{real}). \tag{24}$$

2) Case 2: Fake Samples (Unstable State): For a deepfake sample $\mathbf{q}_{fake}$, the hypothesis posits that it is more likely to lie on a slope with a significant non-zero gradient in the local comparison of interest. Let $\mathbf{g} = \nabla V(\mathbf{q}_{fake})$ be the gradient vector, where $\|\mathbf{g}\| = C \gg 0$.

For a small time step t, we can assume the gradient force is approximately constant (zeroth-order approximation of force). The evolution of momentum is given by:

$$\mathbf{p}(t) = \mathbf{p}(0) + \int_0^t -\nabla V(\mathbf{q}(s))ds \approx -\mathbf{g}t. \tag{25}$$

Under the canonical kinetic energy, $T(t)$ grows quadratically with time:

$$T(t) = \frac{1}{2}\|\mathbf{p}(t)\|^2 = \frac{1}{2}\|\mathbf{g}\|^2 t^2 = \frac{1}{2}C^2 t^2. \tag{26}$$

This confirms the quadratic surge in kinetic energy.

Now, consider the Hamiltonian Action S. Unlike the Lagrangian $(T - V)$, we define S as the accumulation of total energy magnitude $\mathcal{H} = T + V$ to maximize signal detection.

$$S_{fake} = \frac{1}{\tau}\int_0^\tau (T(t) + V(\mathbf{q}(t)))dt. \tag{27}$$

Since $\mathbf{q}_{real}$ is hypothesized to lie near a local minimum while $\mathbf{q}_{fake}$ lies on a higher-gradient slope, we consider the local comparison in which the potential gap $\Delta V = V(\mathbf{q}_{fake}) - V(\mathbf{q}_{real})$ is positive, and $T(t)$ adds a positive quadratic contribution. Specifically, integrating the kinetic term yields:

$$\int_0^\tau T(t)dt \approx \int_0^\tau \frac{1}{2}C^2 t^2 dt = \frac{1}{6}C^2\tau^3. \tag{28}$$

Therefore, the Action gap scales as:

$$S_{fake} - S_{real} \approx \Delta V + \frac{\tau^2}{6}C^2. \tag{29}$$

Comparing Case 1 and Case 2, the positive potential gap together with the additional $\mathcal{O}(\tau^2)$ kinetic term implies:

$$S_{fake} - S_{real} > 0. \tag{30}$$

This completes the proof. □

APPENDIX C
VOLUME PRESERVATION OF THE SYMPLECTIC EULER UPDATE (CONSTANT-MASS CASE)

A desirable property for our HAAD framework is that the measured Energy Dissipation (D) primarily reflects the geometric properties of the data manifold rather than numerical artifacts of the integration scheme. In this section, we show that the **Symplectic Euler** integrator, *in the ideal constant-mass case*, defines a *symplectomorphism*, meaning it strictly preserves the phase-space volume (Liouville's Theorem).

Scope of this result. The analysis below treats $\mathbf{M}$ as a constant (state-independent) positive-definite diagonal matrix. In the main-text implementation (Section III-B2) we instead use a *state-conditioned* preconditioner $\mathbf{M}(\mathbf{q})^{-1}$ in the position update and explicitly omit the $\mathbf{q}$-derivative terms that would appear in an exact variable-mass Hamiltonian flow. Consequently, the constant-mass proof below should be read as motivation for our choice of integrator family (semi-implicit Euler with momentum-before-position ordering) rather than as a strict symplectomorphism guarantee for the implemented preconditioner-augmented update. We accordingly label the implemented rollout "Hamiltonian-inspired, symplectic-style" in Section III-B2 and reserve the strict "symplectomorphism" term for the constant-mass analysis here. We also note upfront that volume preservation does not imply exact Hamiltonian conservation, so D is not fully isolated from numerical effects even in the constant-mass case.

A. Jacobian Analysis of the Discrete Map

Let $\Phi : (\mathbf{q}_t, \mathbf{p}_t) \mapsto (\mathbf{q}_{t+1}, \mathbf{p}_{t+1})$ denote the discrete update map defined by the Symplectic Euler scheme (Equation (9) and (10) in the main text). For a system with d dimensions, the update rule is:

$$\mathbf{p}_{t+1} = \mathbf{p}_t - \eta \nabla V(\mathbf{q}_t), \tag{31}$$

$$\mathbf{q}_{t+1} = \mathbf{q}_t + \eta \mathbf{M}^{-1} \mathbf{p}_{t+1}. \tag{32}$$

To prove volume preservation, we must show that the determinant of the Jacobian matrix of this transformation, $J = \frac{\partial(\mathbf{q}_{t+1}, \mathbf{p}_{t+1})}{\partial(\mathbf{q}_t, \mathbf{p}_t)}$, is identically equal to 1.

The transformation can be decomposed into two sequential sub-steps (shear transformations): i) Momentum Update (Kick): $(\mathbf{q}_t, \mathbf{p}_t) \to (\mathbf{q}_t, \mathbf{p}_{t+1})$ ii) Position Update (Drift): $(\mathbf{q}_t, \mathbf{p}_{t+1}) \to (\mathbf{q}_{t+1}, \mathbf{p}_{t+1})$

Step i): Jacobian of Momentum Update. Differentiating $\mathbf{p}_{t+1}$ with respect to the input state $(\mathbf{q}_t, \mathbf{p}_t)$:

$$J_1 = \begin{pmatrix} \frac{\partial \mathbf{q}_t}{\partial \mathbf{q}_t} & \frac{\partial \mathbf{q}_t}{\partial \mathbf{p}_t} \\ \frac{\partial \mathbf{p}_{t+1}}{\partial \mathbf{q}_t} & \frac{\partial \mathbf{p}_{t+1}}{\partial \mathbf{p}_t} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\eta \mathbf{H}(\mathbf{q}_t) & \mathbf{I} \end{pmatrix}, \tag{33}$$

where $\mathbf{H}(\mathbf{q}_t)$ is the Hessian matrix of potential V. Since this is a lower triangular block matrix with identity blocks on the diagonal, its determinant is:

$$\det(J_1) = \det(\mathbf{I}) \cdot \det(\mathbf{I}) = 1. \tag{34}$$

Step ii): Jacobian of Position Update. Differentiating $\mathbf{q}_{t+1}$ with respect to the intermediate state $(\mathbf{q}_t, \mathbf{p}_{t+1})$:

$$J_2 = \begin{pmatrix} \frac{\partial \mathbf{q}_{t+1}}{\partial \mathbf{q}_t} & \frac{\partial \mathbf{q}_{t+1}}{\partial \mathbf{p}_{t+1}} \\ \frac{\partial \mathbf{p}_{t+1}}{\partial \mathbf{q}_t} & \frac{\partial \mathbf{p}_{t+1}}{\partial \mathbf{p}_{t+1}} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \eta \mathbf{M}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{35}$$

This is an upper triangular block matrix. Similarly, its determinant is:

$$\det(J_2) = 1. \tag{36}$$

Total Jacobian. By the chain rule, the Jacobian of the full step Φ is the product $J = J_2 J_1$. The determinant is:

$$\det(J) = \det(J_2) \det(J_1) = 1 \cdot 1 = 1. \tag{37}$$

B. Physical Implication: Trustworthiness of Dissipation

Since $\det(J) = 1$, the Symplectic Euler integrator preserves the phase-space volume element $d\mathbf{q} \wedge d\mathbf{p}$ exactly.

Remark C.1 (**Reduced Numerical Damping**). Unlike standard Euler or Runge-Kutta methods, where $|\det(J)| \neq 1$ leads to artificial energy drift (numerical damping or explosion), symplectic methods better control long-term energy drift and can often be interpreted as approximately conserving a nearby modified Hamiltonian under suitable step-size assumptions.

This mathematical property means that the symplectic integrator reduces solver-induced energy drift relative to non-symplectic methods. However, volume preservation does not imply exact Hamiltonian conservation, so the observed Energy Dissipation ($D > 0$) is not guaranteed to be exclusively data-induced. The correct interpretation is that D is an **empirically useful discriminative signal**: symplectic integration suppresses some numerical artifacts, making D more attributable to the geometry of the potential surface, but it does not provide a strict isolation guarantee. The empirical effectiveness of D as a detection cue is validated through the ablation study in the main text.

APPENDIX D
ADDITIONAL EXPERIMENTS AND HYPERPARAMETER SENSITIVITY

A. Hyperparameter Sensitivity Analysis

Besides the physical regularization weight λ_{phy} in the main text, we further investigate the sensitivity of the proposed framework to two key hyperparameters: the Hamiltonian evolution step size η, and the energy margin γ. All experiments are conducted on the cross-dataset setting (Train on FF++, Evaluate on Celeb-DF-v2).

i) Impact of Step Size (η). The step size η determines the discrete time interval for the symplectic integrator. Its effect on detection performance is shown in Figure 7.

- **Too Small ($\eta < 0.1$):** The system evolves too slowly. Within the fixed steps $T = 4$, the particle hardly moves from the initial state, resulting in a weak detection signal ($S \approx 0$ for both real and fake).
- **Optimal ($\eta \approx 0.4$):** Provides sufficient 'kick' to reveal the instability of deepfakes while maintaining numerical stability for real samples.
- **Too Large ($\eta > 0.8$):** Leads to numerical instability and gradient explosion, causing the training to diverge.

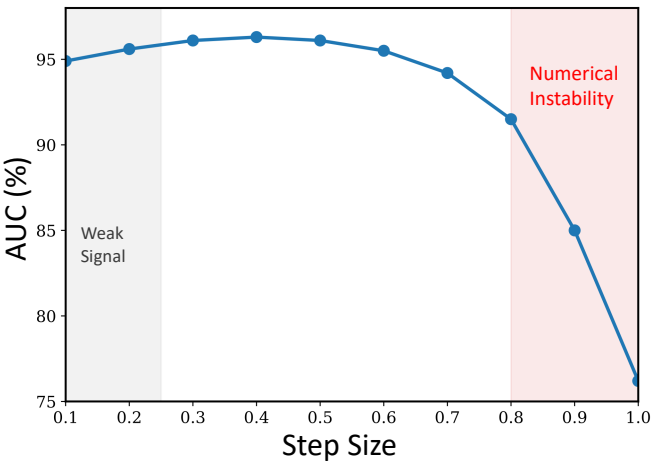

Fig. 7. **Sensitivity to Step Size η.** Performance peaks around $\eta = 0.4$; excessively large steps lead to numerical instability.

B. Solver-Induced vs. Data-Induced Components of D

Because Symplectic Euler conserves a modified (shadow) Hamiltonian with $\mathcal{O}(\eta^2)$ error per step [61], its baseline drift for near-conservative trajectories (real samples) grows predictably with step size η. In contrast, fake-induced drift emerges from large potential gradients and grows faster for larger $\|\nabla V\|$. A complete disentanglement would require measuring real-sample D across multiple η values to isolate the $\mathcal{O}(\eta^2)$ solver baseline, then comparing to fake-sample D at the same η. The integrator ablation in Table VI provides

indirect evidence for this separation: the Euler method, whose $\mathcal{O}(\eta)$ drift raises the solver noise floor, degrades detection performance by blurring the real/fake gap in D, whereas Symplectic Euler's more predictable drift preserves it.

C. Computational Efficiency Analysis

A key concern for physics-based deep learning is the inference latency. In Table XI, we provide a detailed breakdown of the computational cost. The experiments are conducted on a single NVIDIA H100 GPU with a batch size of 32.

TABLE XI
DETAILED COMPARISON OF COMPUTATIONAL COST (INFERENCE PER IMAGE).

Method	Steps T	Time (ms)	Memory (MB)
CLIP ViT-L/14 (backbone)	-	11.5	2450
HAAD (RK4)	4	45.0	3100
HAAD (Euler)	4	12.0	2480
HAAD (Symplectic Euler)	**4**	**14.0**	**2485**

Analysis:

- **Runge-Kutta (RK4):** While precise, it requires 4 gradient evaluations per step, resulting in a latency of **45ms**, which is $\approx 4.3\times$ slower than the backbone. This aligns with the findings in Table 5.
- **Symplectic Euler (Ours):** By requiring only 1 gradient evaluation per step, it achieves an inference time of **14ms**. The slight overhead compared to standard Euler (12ms) stems from the computation of the learnable preconditioner $\mathbf{M}(\mathbf{q})^{-1}$, which provides feature-specific sensitivity.
- **Memory Efficiency:** The Hamiltonian evolution operates in the low-dimensional latent space ($N \times 64$), inducing a negligible memory increase ($\approx$ 35 MB) over the backbone.

APPENDIX E
DETAILED NETWORK ARCHITECTURES AND IMPLEMENTATION

In this section, we provide the specific architectural details of the proposed HAAD framework, including the backbone configuration, the structure of the physical projection heads, the graph topology construction, and the hyperparameter settings.

A. HAAD Module Architecture

The HAAD module projects high-dimensional semantic features into a low-dimensional physical state space. The specific components are defined as follows:

1) Physical State Projection: To reduce computational complexity and enforce a bottleneck, we project the input features $\mathbf{x}$ to a physical state $\mathbf{q} \in \mathbb{R}^{N \times D_{phy}}$ with $D_{phy} = 64$.

$$\mathbf{q} = \text{Linear}(D_{in} \to D_{phy})(\mathbf{x}). \tag{38}$$

2) Mass Estimation Network: The state-conditioned diagonal preconditioner $\mathbf{M}^{-1}(\mathbf{q})$ rescales the position update for each patch and feature dimension. It is estimated via a lightweight MLP with a Softplus activation to keep the scaling positive for numerical stability:

$$\mathbf{M}^{-1}(\mathbf{q}) = \text{Softplus}\left(\text{MLP}(\mathbf{q})\right) + \epsilon, \tag{39}$$

where $\epsilon = 10^{-3}$. The MLP architecture is:

- Layer 1: Linear($64 \to 64$) + ReLU
- Layer 2: Linear($64 \to 1$)

The output is broadcasted to match the dimension of $\mathbf{q}$ for element-wise multiplication.

3) Potential Parameterization Heads: To compute the photometric potential V_{photo}, we estimate intrinsic physical properties using shallow linear heads:

- **Surface Normal (n):** Linear($64 \to 3$) followed by L_2 normalization.
- **Albedo (ρ):** Linear($64 \to 1$) followed by Sigmoid activation.
- **Global Light (l):** Linear($64 \to 3$), averaged over all patches to obtain a global vector.

B. Graph Topology Construction

To compute the geometric potential V_{geo}, we construct a spatial graph $\mathcal{G}$ over the patch grid. Unlike fully connected graphs, we utilize a local k-NN graph to capture local structural continuity.

- **Coordinates:** We assign 2D grid coordinates (y_i, x_i) to each patch.
- **Connectivity:** For each patch, we connect it to its $k = 8$ nearest spatial neighbors.
- **Weights:** The edge weights w_{ij} are computed using a Gaussian kernel based on Euclidean distance d_{ij}:

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), \quad \text{with } \sigma = 8.0. \tag{40}$$

- **Laplacian:** We compute the sparse Graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the degree matrix.

C. Symplectic Integration Scheme

We utilize the Symplectic Euler method for the discrete evolution of the system. Let η be the step size and T be the number of steps. The update rule at step t is implemented as:

$$\mathbf{F}_t = -\nabla_{\mathbf{q}} V(\mathbf{q}_t), \tag{41}$$

$$\mathbf{p}_{t+1} = \mathbf{p}_t + \eta \cdot \mathbf{F}_t, \tag{42}$$

$$\mathbf{q}_{t+1} = \mathbf{q}_t + \eta \cdot (\mathbf{M}^{-1} \odot \mathbf{p}_{t+1}). \tag{43}$$

This implementation is the same semi-implicit momentum-before-position update used in the main text, with $\mathbf{F}_t = -\nabla_{\mathbf{q}} V(\mathbf{q}_t)$ and the state-conditioned preconditioner applied only in the position update.