When Task-Optimal RF Signals Are Physically Invalid: Physics-Guided Generation for Resolving the Task–Physics Conflict

***Abstract*—Generative models for RF signal synthesis are typically optimized for task utility, implicitly assuming that task-optimal samples are also physically valid. We empirically invalidate this assumption: across three scene configurations, samples optimized for high task confidence consistently violate physical constraints, whereas physically valid samples rarely achieve peak task utility. This task–physics conflict persists across environments and cannot be resolved by post-hoc filtering or larger sample budgets.**

We introduce ConflictRate, a diagnostic metric that quantifies the fraction of task-confident samples violating physical constraints. Using ConflictRate, we demonstrate that post-hoc quality-guided selection (e.g., SynCheck) exacerbates the issue, raising the conflict rate from 72% (unguided baseline) to 85% by concentrating temporally inconsistent samples. In contrast, physics-guided generation (AetherGen) drives the conflict rate to 0% by enforcing coherent delay evolution across time. Independent validation across three domains—path-level smoothness, channel impulse response (CIR) temporal stability, and channel state information (CSI) phase coherence—confirms that this improvement is not an artifact of the evaluation metric.

A mechanistic analysis identifies temporal inconsistency ($\mathcal{L}_{\text{motion}}$) as the dominant failure mode. We show that resolving this single inconsistency is sufficient to eliminate the mismatch, establishing that the task–physics conflict is a fundamental formulation gap, not a limitation of model capacity or data scale.

This work suggests a shift in generative RF modeling: enforcing temporal physical structure at generation time may be more fundamental than scaling task objectives or enlarging candidate pools.

***Index Terms*—RF sensing, generative model, diffusion model, physics-guided sampling, multipath channel, digital twin, wireless channel simulation, synthetic data generation**

I. Introduction

Generative models are increasingly used to synthesize RF sensing data for simulation, data augmentation, and downstream learning tasks. Recent advances in diffusion-based models [1], [2] have enabled high-fidelity generation of complex signal structures, making synthetic RF data a promising alternative to costly real-world measurements. In many applications, such as gesture recognition, occupancy sensing, and wireless channel simulation, generative models are typically evaluated based on their utility for downstream tasks, e.g., classification accuracy or task-specific confidence.

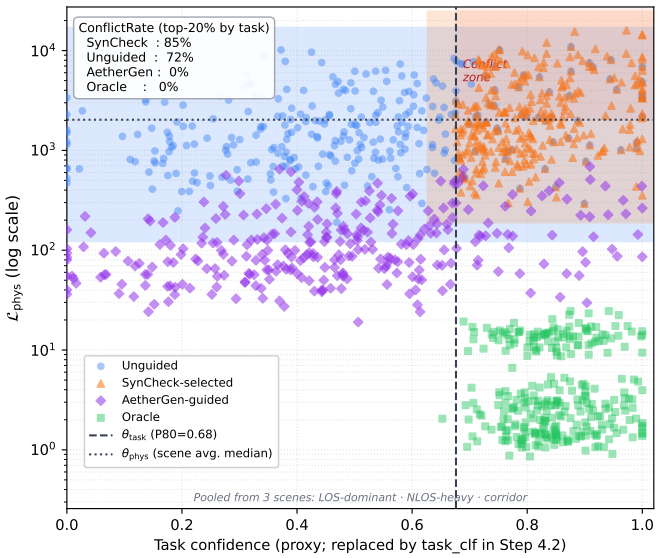

Fig. 1. **Task–physics mismatch in RF generative models.** Scatter plot of task confidence $U_{\text{task}}(x)$ versus physical violation $\mathcal{L}_{\text{phys}}(x)$ for generated samples. High-confidence samples frequently exhibit large physical violations, while physically valid samples are not necessarily task-optimal. The shaded region indicates the conflict zone used to compute *ConflictRate*. This result demonstrates that optimizing for task utility alone does not guarantee physical validity.

However, optimizing for task utility alone does not guarantee that generated signals are physically valid. In RF sensing, signal structure is governed by propagation physics—hereafter defined as structural constraints—including multipath geometry, energy conservation, and temporal coherence. A model may produce samples that appear useful for a task while violating these physical constraints. This raises a fundamental question:

Do task-optimized generative models produce physically valid RF signals?

In this paper, we show that the answer is often no. Through a systematic empirical study, we identify a consistent mismatch between task utility and physical validity, which we term the **task–physics conflict**. As illustrated in Fig. 1, samples with high task confidence frequently exhibit large physical violations, while physically valid samples do not necessarily achieve high task utility. We quantify this phenomenon using a conflict rate metric and demonstrate that the conflict persists across multiple

environments and selection thresholds (Fig. 2).

A natural approach to improving generation quality is post-hoc filtering, where a large pool of generated samples is ranked by a task score and only the top candidates are selected. However, we show that this strategy exacerbates the problem: selection based on task utility systematically concentrates samples in regions of high physical violation (Fig. 3). This reveals a fundamental limitation of selection-based pipelines, where improving task performance can come at the expense of physical validity.

To understand the origin of this conflict, we perform a component-wise analysis of physical consistency and find that **temporal inconsistency is the dominant failure mode** in unconstrained generation. Specifically, generated path delays exhibit unstable temporal evolution, which leads to large violations in physical validity metrics. This observation motivates a shift from selection-based methods to *generation-time constraint enforcement*.

We therefore propose **AetherGen**, a physics-guided generative framework that integrates domain-specific physical constraints directly into the sampling process of a diffusion model. Instead of selecting valid samples after generation, AetherGen steers the generative trajectory toward physically consistent regions by applying gradient-based corrections derived from a composite physics loss. This enables the model to produce samples that are both task-relevant and physically meaningful.

Extensive experiments demonstrate that AetherGen significantly reduces physical violations compared to both unguided generation and post-hoc filtering, while maintaining comparable task utility. Importantly, we verify that these improvements are not an artifact of metric design: reductions in the physics loss correlate strongly with independent physical measures that are not used during optimization (Fig. 5). Furthermore, we show that the task–physics conflict is not due to insufficient model capacity or limited data. Even with stronger task models or larger sample pools, the conflict persists (Fig. 7), indicating that it is a *formulation-level issue* rather than a modeling limitation.

Our findings suggest a broader perspective on generative modeling for physical systems: optimizing for downstream task performance alone is insufficient when physical consistency is required. Instead, incorporating structured domain constraints during generation provides a more principled and effective solution. While our study focuses on RF sensing, the proposed framework and insights may extend to other domains where physical validity plays a critical role, such as robotics, simulation, and scientific data generation.

a) Contributions.: The main contributions of this work are as follows:

- We identify and formalize the **task–physics conflict** in RF generative modeling, showing that high task utility does not imply physical validity.
- We demonstrate that this conflict is **robust across environments and selection thresholds**, and that post-hoc filtering exacerbates the problem.
- We provide a **mechanistic explanation** by showing that temporal inconsistency is the dominant source of physical violation.
- We propose **AetherGen**, a physics-guided diffusion framework that enforces physical constraints during generation.
- We validate that improvements are **physically meaningful and not metric-specific**, and show that the problem is not resolved by stronger models or more data.

II. Related Work

A. Synthetic Data for RF Sensing

Synthetic data generation has become an important tool in RF sensing, enabling scalable data augmentation and simulation without extensive measurement campaigns. Recent works such as RF Genesis [3] propose generative models for synthesizing RF signals directly from learned distributions.

However, these approaches primarily optimize for task performance or distributional similarity, without explicitly enforcing physical consistency. As a result, generated samples may exhibit unrealistic multipath structures or temporal dynamics despite achieving high task accuracy.

B. Quality-Guided Utilization of Synthetic Data

A complementary line of work focuses on improving synthetic data quality through post-hoc selection. For example, SynCheck [4] proposes a quality-guided framework where samples are selected based on task affinity.

While effective in improving downstream performance, such approaches implicitly assume that task-optimal samples are also physically valid. Our results challenge this assumption by showing that task-driven selection can favor physically inconsistent samples.

C. Diffusion Models and Guided Sampling

Diffusion models [1], [2] have achieved strong performance in generative modeling. Guided sampling techniques such as classifier guidance [5] and score-based guidance [6] steer generation toward desired outputs.

Our work differs in that we introduce physics-guided sampling, which enforces structured constraints rather than task labels, and we analyze its role in resolving task–physics mismatch.

D. Physics-Informed Learning

Physics-informed approaches such as PINNs [7] embed physical constraints into model training. These methods ensure that learned models satisfy governing equations.

In contrast, our work focuses on enforcing physical consistency during generation rather than training, highlighting a complementary perspective on physics-aware modeling.

E. Wireless Channel Modeling

Classical wireless channel models characterize multipath propagation through stochastic or geometry-based formulations. Stochastic models such as the Saleh–Valenzuela model [8] describe cluster-structured multipath arrivals, while standardized models including 3GPP TR 38.901 [9] and COST 2100 [10] extend these to geometry-based stochastic channels (GBSM) that capture spatial and temporal correlation. These models encode physical priors (delay spread, angular spread, Doppler) through distributional constraints derived from measurement campaigns.

Our work is complementary: rather than fitting statistical parameters to measurements, we enforce similar physical constraints as differentiable loss terms that guide a learned generative model at inference time. This allows physical consistency to be maintained without requiring explicit geometry or measurement-derived parameterization.

F. Ray Tracing and Digital Twins

Ray-tracing-based simulation provides physically accurate RF channel data by modeling electromagnetic wave propagation through detailed 3D scene geometry. DeepMIMO [11] generates large-scale MIMO channel datasets from ray tracing, enabling geometry-aware evaluation of wireless algorithms. Sionna RT [12] introduces differentiable ray tracing that integrates with deep learning pipelines, supporting gradient-based optimization over scene parameters. More broadly, digital twin frameworks [13] construct virtual RF environments to replicate physical deployments and support network planning.

While these methods achieve high physical fidelity, they require detailed scene geometry and incur substantial computational cost, limiting scalability across diverse deployment conditions. Our approach instead uses a learned generative model with physics-guided sampling, achieving physical consistency without scene geometry and scaling efficiently to new environments.

III. Problem Formulation

A. Path-level RF Representation

We consider RF signals represented at the multipath level, where each sample is a time-evolving set of propagation paths:

$$x \in \mathbb{R}^{T \times L \times C}, \tag{1}$$

with T time frames, L paths, and channel attributes C (e.g., delay τ, amplitude α).

This representation explicitly encodes both geometric structure and temporal dynamics, which are central to physical validity.

B. Task Utility vs. Physical Validity

Let $s_{\text{task}}(x) \in [0, 1]$ denote the scalar utility score of a sample with respect to a downstream task (e.g., classification confidence).

Independently, we define a physics-based loss:

$$\mathcal{L}_{\text{phys}}(x) = \sum_k \mathcal{L}_k(x), \tag{2}$$

where each $\mathcal{L}_k$ encodes a domain-specific constraint (e.g., delay structure, energy consistency, temporal coherence).

Lower $\mathcal{L}_{\text{phys}}$ indicates higher physical validity.

a) Key observation.: Task utility and physical validity are *not inherently aligned*. In particular, high $s_{\text{task}}(x)$ does not guarantee low $\mathcal{L}_{\text{phys}}(x)$.

C. Task–Physics Conflict

We formalize this misalignment as a *task–physics conflict*.

a) Oracle samples.: We define *oracle samples* as path realizations generated from a trusted physics-based process that reflects ground-truth RF propagation behavior. In our setting, oracle samples are obtained from the AetherSim simulator under known scene configurations, without task-driven selection or generative bias. Oracle samples define a reference distribution of physically consistent path realizations; they are not used for sample selection or individual-sample optimization.

Let θ_{task} denote a high-percentile threshold (e.g., top-k% by task score), and θ_{phys} a reference physical threshold. θ_{phys} is defined relative to oracle statistics, providing a consistent reference for physical validity. In practice, θ_{phys} is set at the **75th percentile** of $\mathcal{L}_{\text{phys}}$ evaluated on oracle samples, classifying a generated sample as physically invalid only if its physics loss exceeds three-quarters of the physically realizable population.

A sample exhibits conflict if:

$$s_{\text{task}}(x) > \theta_{\text{task}} \quad \text{and} \quad \mathcal{L}_{\text{phys}}(x) > \theta_{\text{phys}}. \tag{3}$$

We define the *ConflictRate* as the fraction of such samples within the top task-selected subset. ConflictRate is not merely a diagnostic indicator; it is the expected fraction of task-optimized samples whose temporal structure lies outside the Doppler-constrained physical manifold—a deterministic consequence of objective-level structural invariance, independent of model capacity or data scale.

D. Failure of Selection-Based Pipelines

A widely adopted strategy in practice is *post-hoc selection*, where samples are generated and filtered based on task-related scores.

This implicitly assumes:

$$s_{\text{task}}(x) \uparrow \Rightarrow \ \mathcal{L}_{\text{phys}}(x) \downarrow, \tag{4}$$

i.e., task utility serves as a proxy for physical validity.

As we will show empirically (Sec. VI), this assumption does not hold, leading to systematic selection of physically invalid samples.

E. Objective

Given a generative model $p_\theta(x)$, the goal is to produce samples that are both:

$$\text{(i) high task utility} \quad \text{and} \quad \text{(ii) physically valid.} \tag{5}$$

This leads to the central question of this work:

> *How can we enforce physical validity during generation, rather than relying on post-hoc selection?*

We address this by introducing a physics-guided generation framework that directly optimizes for physical consistency during sampling.

For clarity, throughout this paper we use *oracle* to denote physics-consistent reference samples, *reference* for derived statistics used in loss computation, and *baseline* for the comparison methods (Unguided, SynCheck).

IV. Empirical Evidence of Task–Physics Conflict

We begin with a guiding hypothesis:

> *Task-driven selection induces a bias toward samples that are optimal for the task objective but violate structured physical constraints as the volume of physically invalid states within a task-equivalent class typically exceeds that of valid ones.*

Formally, this suggests that:

$$\arg\max_x s_{\text{task}}(x) \not\subseteq \arg\min_x \mathcal{L}_{\text{phys}}(x), \tag{6}$$

i.e., task-optimal samples are not guaranteed to be physically valid.

a) Proposition (Selection Bias under Invariant Objectives).: Let $s_{\text{task}}(x)$ be a task objective that is invariant to a set of structured transformations $\mathcal{T}$ (e.g., temporal permutations or phase-consistent variations), while $\mathcal{L}_{\text{phys}}(x)$ is sensitive to these transformations. Then, selecting samples by maximizing $s_{\text{task}}(x)$ induces a bias toward configurations that violate the structural constraints encoded by $\mathcal{L}_{\text{phys}}$.

Proof 1: Let $\mathcal{T}$ be the set of structured transformations to which s_{task} is invariant by assumption. For any sample x, the equivalence class $\mathcal{T}(x) = \{x' : s_{\text{task}}(x') = s_{\text{task}}(x)\}$ contains all orbits of $\mathcal{T}$ applied to x. Since $\mathcal{L}_{\text{phys}}$ is sensitive to $\mathcal{T}$, there exist members $x, x' \in \mathcal{T}(x)$ with $\mathcal{L}_{\text{phys}}(x') > \mathcal{L}_{\text{phys}}(x)$. Selection by s_{task} assigns equal probability to all members of $\mathcal{T}(x)$; therefore the expected $\mathcal{L}_{\text{phys}}$ of selected samples strictly exceeds that of the physically optimal member within the class. □

b) Interpretation.: Task models often operate on aggregated or order-invariant features, and therefore cannot distinguish between physically consistent and inconsistent realizations within the same equivalence class. As a result, optimization or selection based solely on task utility may favor samples that satisfy the task objective while violating structured dependencies such as temporal coherence.

c) Implication (Corollary of Proposition 1).: Under the above conditions, increasing model capacity or sample size does not eliminate the task–physics conflict, as the bias arises from the objective itself rather than from model expressiveness.

In the following, we empirically validate this hypothesis by analyzing the relationship between task utility and physical validity.

A. Conflict in Task–Physics Space

Fig. 1 visualizes samples in the joint space of task utility $s_{\text{task}}(x)$ and physical violation $\mathcal{L}_{\text{phys}}(x)$.

We observe a substantial region where samples with high task confidence exhibit large physical violations. In particular, many samples in the top percentile of task scores lie far from the physically valid region.

This directly contradicts the implicit assumption that task-optimal samples are physically plausible.

a) Key finding.: High task utility does not imply physical validity.

To quantify this phenomenon, we measure the ConflictRate (Sec. III), defined as the fraction of high-confidence samples that violate physical constraints. We find that a significant portion of top-ranked samples fall into the conflict region.

B. Robustness Across Environments and Thresholds

A natural concern is whether the observed mismatch is an artifact of a particular dataset or threshold choice.

To address this, Fig. 2 evaluates ConflictRate across multiple scene types and selection thresholds.

Across all environments—including LOS-dominant, NLOS-heavy, and cluttered scenarios—the ConflictRate remains consistently high. Similarly, varying the task threshold from moderate (P60) to extreme (P95) does not eliminate the conflict.

a) Key finding.: The task–physics mismatch is robust and not tied to a specific setting.

Notably, ConflictRate remains near-constant across the full threshold range (P60–P95), exhibiting step-like threshold-invariant behavior rather than a gradual transition. This is a signature of the objective-level formulation gap established in Proposition 1: once the task objective becomes invariant to temporal physical structure, the conflict is entered sharply and persists regardless of the operating point.

This suggests that the misalignment arises from the structure of the problem itself, rather than from dataset bias or threshold selection.

C. Failure of Post-hoc Selection

We next evaluate whether post-hoc selection can resolve the mismatch.

Fig. 3(a) shows that filtering samples based on task scores improves task utility but increases physical violations. In other words, selection operates along a trade-off frontier rather than resolving the underlying conflict.

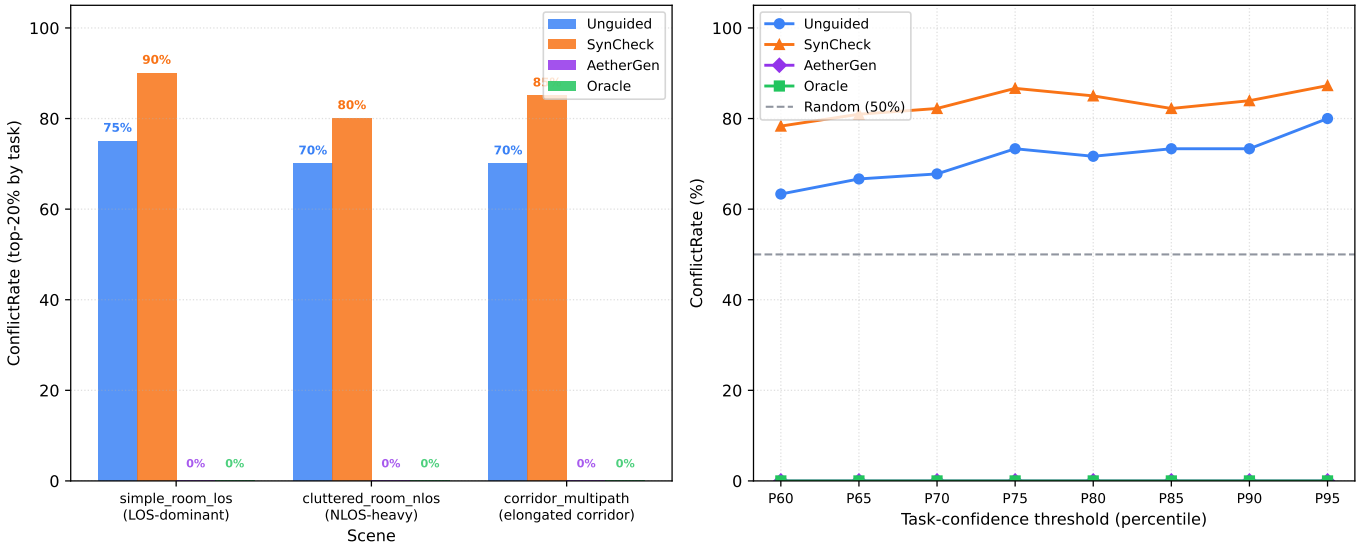

Fig. 2. **Robustness of the task–physics mismatch.** (a) ConflictRate across different scene types (LOS-dominant, NLOS-heavy, and corridor environments). (b) ConflictRate under varying task-confidence thresholds (P60–P95). The mismatch persists across environments and thresholds, indicating that it is not an artifact of a specific dataset or evaluation setting.

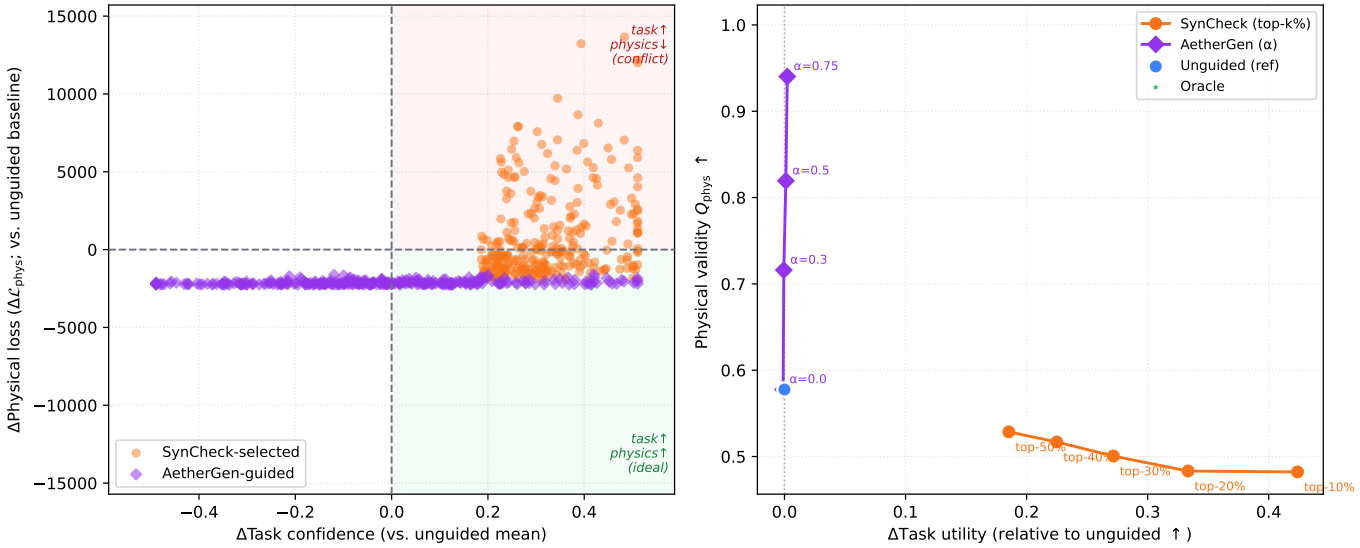

Fig. 3. **Comparison of post-hoc selection and physics-guided generation.** (a) Change in task utility and physical violation relative to unguided generation. Post-hoc selection (SynCheck) improves task utility but increases physical violations, indicating a trade-off. (b) Pareto view of task utility vs. $\mathcal{L}_{\text{phys}}$. AetherGen shifts samples toward lower physical violation without sacrificing task performance, while filtering moves along the trade-off frontier.

This behavior is further illustrated in Fig. 3(b), where filtering moves samples toward higher task confidence at the expense of increased $\mathcal{L}_{\text{phys}}$.

a) Key finding.: Selection improves task metrics at the cost of physical validity.

Thus, post-hoc filtering cannot bridge the task–physics gap, as it does not alter the generative process itself.

In the next section, we introduce a physics-guided generation framework that addresses this limitation by incorporating domain constraints directly into the sampling process.

V. AetherGen: Physics-Guided Generation

We propose **AetherGen**, a physics-guided generative framework that enforces physical consistency during the sampling process of a diffusion model. Unlike post-hoc filtering, which selects plausible samples after generation, AetherGen integrates domain-specific physical constraints directly into the generative dynamics, steering samples toward physically valid regions.

A. Preliminaries: Diffusion-Based Generation

We consider a diffusion model that generates RF path representations $x \in \mathbb{R}^{T\times L\times C}$, where T denotes the number of time frames, L the number of paths, and C the set of channel attributes (e.g., delay τ, amplitude α, and coupling c).

Starting from Gaussian noise $x_T \sim \mathcal{N}(0, I)$, the model iteratively denoises the sample through a reverse process:

$$x_{t-1} = f_\theta(x_t, t), \tag{7}$$

where f_θ is a learned denoising function (the ε-predictor parameterized by θ).

In this work, we adopt a simple MLP-based backbone for f_θ that processes flattened path representations with a timestep embedding. This design intentionally avoids strong temporal inductive biases, allowing us to isolate the effect of physics guidance.

B. Physics-Based Metrics

To evaluate physical validity, we define a composite metric $\mathcal{L}_{\text{phys}}$ that captures structured consistency in path-level RF signals. This metric is used for evaluation and for defining physics-guided corrections during sampling, but is not a task objective.

We distinguish between two types of metrics: (i) physically-derived constraints reflecting fundamental properties of RF propagation, and (ii) statistical proxies capturing structured dependencies when explicit geometry is not directly observable.

a) Physically-derived metrics.: **Delay consistency.** We measure discrepancy between delay distributions using a Wasserstein-1 distance:

$$\mathcal{L}_{\text{delay}} = \mathbb{E}\left[\int |F_{\text{gen}}(\tau) - F_{\text{ref}}(\tau)|\, d\tau\right]. \tag{8}$$

Wasserstein-1 is adopted due to its robustness to permutation and path-order ambiguity.

Energy consistency. We enforce approximate conservation of path energy:

$$\mathcal{L}_{\text{energy}} = \mathbb{E}\left[\frac{\left|\sum_\ell |\alpha_\ell^{\text{gen}}|^2 - \sum_\ell |\alpha_\ell^{\text{ref}}|^2\right|}{\sum_\ell |\alpha_\ell^{\text{ref}}|^2}\right]. \tag{9}$$

While explicit geometric constraints such as $\tau = d/c$ are not directly observable in our representation, these distributional and energy-based measures provide a physically grounded approximation.

b) Statistical proxy metrics.: **Coupling consistency.** Here $c_\ell \in [0, 1]$ denotes the relative power fraction (coupling efficiency) of path ℓ, encoding how each propagation component partitions energy at the receiver. Physically valid signals exhibit geometry-consistent power distributions across paths; $\mathcal{L}_{\text{coupling}}$ penalizes deviations from this reference distribution. We constrain the mean coupling ratio:

$$\mathcal{L}_{\text{coupling}} = \mathbb{E}\left[\left|\bar{c}^{\text{gen}} - \bar{c}^{\text{ref}}\right|\right], \quad \bar{c} = \frac{1}{L}\sum_{\ell=1}^{L} c_\ell. \tag{10}$$

Temporal consistency. We measure mismatch in temporal variation:

$$\mathcal{L}_{\text{motion}} = \mathbb{E}\left[\frac{\left|\text{Var}_t(\tau^{\text{gen}}) - \text{Var}_t(\tau^{\text{ref}})\right|}{\text{Var}_t(\tau^{\text{ref}})}\right]. \tag{11}$$

In indoor multipath propagation, path delays are bounded by the Doppler constraint $|d\tau_\ell/dt| \le 2v_{\max}/c$, giving a physical ceiling of approximately 13 ns/s for $v_{\max}$=2 m/s at sub-6 GHz. The oracle statistics implicitly encode this Doppler-constrained manifold; variance is adopted as a distribution-level proxy that captures violations without requiring explicit per-frame velocity estimation.

c) Composite metric.: The overall physics metric is:

$$\mathcal{L}_{\text{phys}} = \mathcal{L}_{\text{delay}} + \mathcal{L}_{\text{energy}} + \mathcal{L}_{\text{coupling}} + \mathcal{L}_{\text{motion}}. \tag{12}$$

d) Reference construction.: Reference statistics are estimated from independent oracle samples and are not used for sample selection. We match distributions rather than individual samples, preventing instance-level regression toward specific oracle realizations.

e) Units and normalization.: All terms are normalized except $\mathcal{L}_{\text{delay}}$, which is measured in delay units. This ensures scale-invariant comparisons across scenes.

f) Training vs. evaluation.: These metrics are not optimized directly during model training, but are used for analysis and for defining physics-guided corrections during sampling.

C. Physics-Guided Sampling

To resolve the task–physics mismatch, we incorporate physics constraints directly into the diffusion sampling process. Unlike classifier or score guidance, which relies on learned labels or likelihoods, our approach operates on structured physical constraints defined in the RF signal domain.

a) Guided reverse process.: Given a diffusion model with denoising function f_θ, we modify the reverse update as:

$$x_{t-1} = f_\theta(x_t, t) - \lambda \nabla_{x_t} \mathcal{L}_{\text{phys}}(x_t), \tag{13}$$

where $\lambda \ge 0$ controls the strength of physics guidance.

b) Interpretation.: The first term follows the learned data distribution, while the second term applies a structured correction that penalizes violations of physical consistency. This reshapes the sampling trajectory toward regions that satisfy both task utility and physical validity.

c) Relation to existing guidance.: Existing guidance methods bias generation using learned objectives defined in label or feature space. In contrast, our guidance operates directly in the signal domain using physically interpretable constraints. This allows enforcement of propagation-consistent structure rather than semantic alignment.

d) Energy-based perspective.: This formulation can be viewed as energy-based guidance, where the energy function is defined by $\mathcal{L}_{\text{phys}}$. The sampler therefore favors low-energy (physically consistent) configurations during generation.

e) Effect on task–physics alignment.: By injecting gradients of $\mathcal{L}_{\text{phys}}$ at each sampling step, the model selects physically valid representatives within task-equivalent regions. This reduces temporal instability and energy inconsistency without relying on large candidate pools or post-hoc filtering.

f) Practical considerations.: In practice, gradients are computed with respect to x_t at each timestep. We use a fixed guidance scale $\lambda = 1.0$, which provides a stable balance between constraint enforcement and sample diversity. The physics loss is applied only during sampling-time guidance, and does not alter the training objective.

D. Relation to Post-hoc Filtering

Post-hoc filtering selects the best samples from a large candidate pool based on a task or physics score. In contrast, AetherGen reduces the need for large candidate sets by *guiding the generation process itself.*

This distinction is critical:

- Filtering improves sample quality through *selection.*
- AetherGen improves sample quality through *generation.*

As shown in Sec. VI, guidance achieves lower physical violation than filtering, even when using significantly fewer samples.

E. Implementation Details

We use a deterministic DDIM sampler with $T = 50$ steps. The guidance scale λ is fixed to 1.0 in all experiments, which provides a stable balance between constraint enforcement and sample diversity.

Gradients of $\mathcal{L}_{\text{phys}}$ are computed with respect to the current sample x_t and applied at each sampling step. All comparisons between guided and unguided generation use identical initial noise and matched oracle samples to ensure fairness.

VI. Experimental Evaluation

We evaluate AetherGen to answer three key questions: (1) Does a task–physics mismatch exist in RF generative models? (2) Can post-hoc selection resolve this mismatch? (3) Can physics-guided generation improve physical validity without sacrificing task utility?

A. Experimental Setup

a) Data generation.: We use a synthetic RF path dataset where each sample is represented as $x \in \mathbb{R}^{T \times L \times C}$, with $T = 16$ frames, $L = 8$ paths, and channel attributes including delay, amplitude, and coupling.

To ensure generality, scenes are sampled from a parameterized distribution covering three environment types: (1) LOS-dominant (simple room), (2) NLOS-heavy (cluttered room), and (3) multipath-rich (corridor). This prevents overfitting to a single deterministic environment. All results are aggregated across environments.

b) Signal conditions.: To ensure realistic signal conditions, we apply measurement noise comprising (i) delay jitter ($\sigma_\tau = 0.25\,\text{ns}$), (ii) amplitude noise ($\sigma_a = 0.5\,\text{dB}$ per path), and (iii) phase noise consistent with oscillator impairments ($\sigma_\phi = 3^\circ$ RMS per frame).

c) Compared methods.: We compare four conditions:

- **Unguided**: diffusion model without physics guidance.
- **SynCheck**: post-hoc selection based on task confidence.
- **AetherGen**: physics-guided generation (ours).
- **Oracle**: physics-consistent reference samples obtained from AetherSim (defined in Sec. III).

d) Metrics.: Task utility is measured using a confidence proxy $U_{\text{task}}(x)$. Physical validity is measured using $\mathcal{L}_{\text{phys}}$ (Sec. V).

ConflictRate is computed as defined in Sec. III, with θ_{phys} set at the 75th percentile of $\mathcal{L}_{\text{phys}}$ evaluated on oracle samples.

B. Task–Physics Conflict

Fig. 1 plots task confidence against $\mathcal{L}_{\text{phys}}$. We observe a clear mismatch: high-confidence samples frequently exhibit large physical violations, while physically valid samples are not necessarily task-optimal.

The ConflictRate exceeds 70% for unguided generation, indicating that optimizing for task utility alone does not ensure physical consistency.

C. Robustness Across Environments and Thresholds

We test whether this mismatch is scenario-dependent.

Fig. 2(a) shows ConflictRate across LOS-dominant, NLOS-heavy, and corridor scenes. As quantified in Fig. 2(a), the Conflict Rate remains consistently above 70% across LOS-dominant, NLOS-heavy, and corridor scenes.

Fig. 2(b) varies the task threshold from P60 to P95. ConflictRate remains consistently high, demonstrating that the phenomenon is not sensitive to threshold selection.

Conclusion: The task–physics mismatch is robust and not tied to specific datasets or evaluation settings.

D. Failure of Post-hoc Selection

We evaluate whether filtering can resolve the mismatch.

Fig. 3(a) shows that SynCheck improves task utility but increases $\mathcal{L}_{\text{phys}}$, indicating a trade-off.

Fig. 3(b) presents a Pareto view: filtering shifts samples toward higher task confidence but lower physical validity.

Conclusion: Post-hoc selection cannot resolve the mismatch and may exacerbate physical violations.

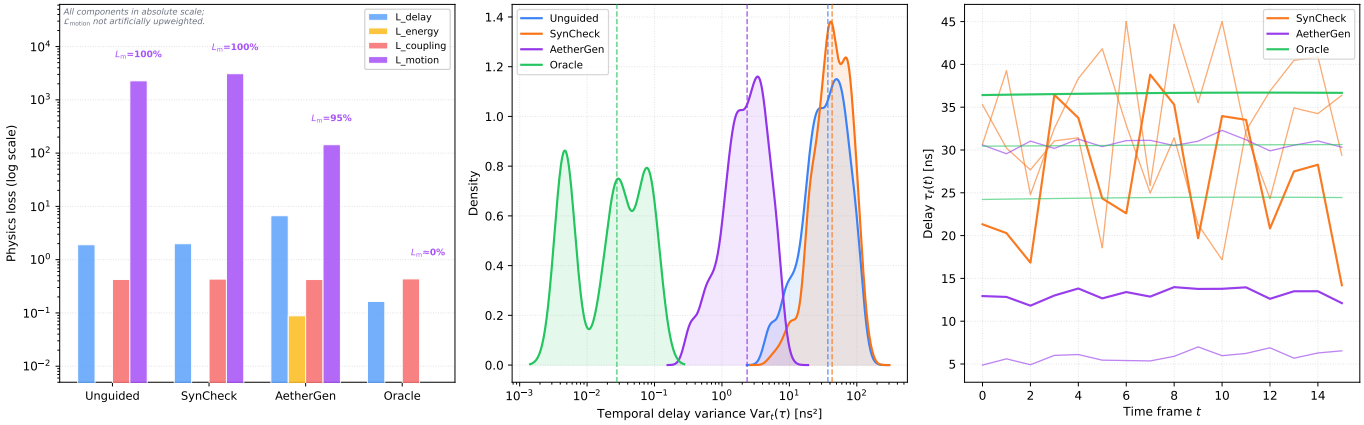

Fig. 4. **Mechanistic analysis of the task–physics mismatch.** (a) Decomposition of $\mathcal{L}_{\text{phys}}$ shows that temporal inconsistency ($\mathcal{L}_{\text{motion}}$) dominates physical violations. (b) Distribution of temporal variance $\mathrm{Var}_t(\tau)$, showing that AetherGen reduces variance toward the oracle distribution. (c) Example trajectories of path delay over time. Unguided and task-selected samples exhibit unstable temporal behavior, while AetherGen produces smooth and physically consistent evolution.

E. Effectiveness of Physics-Guided Generation

We evaluate AetherGen under the same conditions.

As shown in Fig. 3, AetherGen significantly reduces physical violations while maintaining comparable task utility. Unlike filtering, it does not introduce a trade-off.

Across all environments, AetherGen reduces ConflictRate to near zero, indicating that high task confidence is no longer associated with physical inconsistency. This improvement is qualitatively evident in Fig. 4(c), which visualizes the temporal evolution of path delays. While unguided and task-selected models produces jittery and physically improbable trajectories, AetherGen maintains smooth, coherent path evolution that closely aligns with ground-truth patterns.

F. Independent Validation Beyond $\mathcal{L}_{phys}$

To verify that improvements are not due to optimizing a specific metric, we evaluate independent physical observables across three domains, none of which share components with $\mathcal{L}_{\text{phys}}$.

a) Path-domain validation.: We measure an independent temporal smoothness score $1/(1+\overline{|\Delta\tau_t|})$, which depends on first-order frame-to-frame delay changes rather than the temporal variance term used in $\mathcal{L}_{\text{motion}}$. Although both metrics depend on delay evolution, they are computed from distinct formulations and share no components. Fig. 5(a) shows a strong negative correlation between $\mathcal{L}_{\text{phys}}$ and this score (Pearson $r{=}-0.92$, $p{<}0.001$), indicating that lower physics loss corresponds to smoother physical evolution. Fig. 5(b) confirms that AetherGen improves delay smoothness over both Unguided and SynCheck, while post-hoc selection alone does not yield comparable gains.

b) Channel impulse response (CIR)-level validation.: As a second independent check, we compute the RMS delay spread per frame, $\sigma_\tau(t) = \sqrt{\sum_\ell p_\ell(\tau_\ell - \bar{\tau})^2}$, where $p_\ell = |\alpha_\ell|^2 / \sum_\ell |\alpha_\ell|^2$, and measure its temporal coefficient of variation $\mathrm{CoV}_t(\sigma_\tau)$. This metric has no overlap with any term in $\mathcal{L}_{\text{phys}}$ and captures whether generated paths induce stable channel impulse responses over time. Fig. 5(c) shows that AetherGen achieves a mean $\mathrm{CoV}_t{=}0.043$, a 3.9× reduction compared to Unguided (0.166) and SynCheck (0.182). Oracle channels achieve near-zero variation (0.003), confirming physical stability; the full distributions and KS-test statistics are provided in Fig. 6.

c) Channel state information (CSI)-level validation.: As a third check, we evaluate *temporal CSI coherence* — the average normalized inner product between consecutive synthesized CSI frames:

$$\rho(n) = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\left|\langle H_n(t{+}1),\, H_n(t)\rangle\right|}{\|H_n(t{+}1)\|\,\|H_n(t)\|}, \tag{14}$$

where $H_n(t) \in \mathbb{C}^K$ is synthesized via $H(k,t) = \sum_\ell \alpha_\ell(t)\exp(-j2\pi f_k \tau_\ell(t))$ with $f_c{=}6\,\text{GHz}$, $B{=}80\,\text{MHz}$, $K{=}64$, consistent with 802.11ax (Wi-Fi 6) 80 MHz operation [14]. This metric is constructed entirely in the frequency domain and shares no terms with $\mathcal{L}_{\text{phys}}$. Fig. 5(d) shows that AetherGen achieves $\rho{=}0.495$, a +12% improvement over Unguided ($\rho{=}0.442$, $p{<}0.001$, $N{=}300$). The smaller gain relative to the CIR domain is expected, as CSI is an indirect frequency-domain projection of path-level structure.

Together, all three domains yield consistent ordering: Oracle $\gg$ AetherGen $\gg$ Unguided $\approx$ SynCheck, providing cross-domain evidence that physics-guided generation improves real physical consistency rather than merely minimizing $\mathcal{L}_{\text{phys}}$.

G. Is the Mismatch a Model or Data Issue?

We test whether the mismatch can be resolved by stronger models or larger sample sizes.

Fig. 7(a) increases task model strength. ConflictRate increases rather than decreases, indicating that better task models do not resolve the issue.

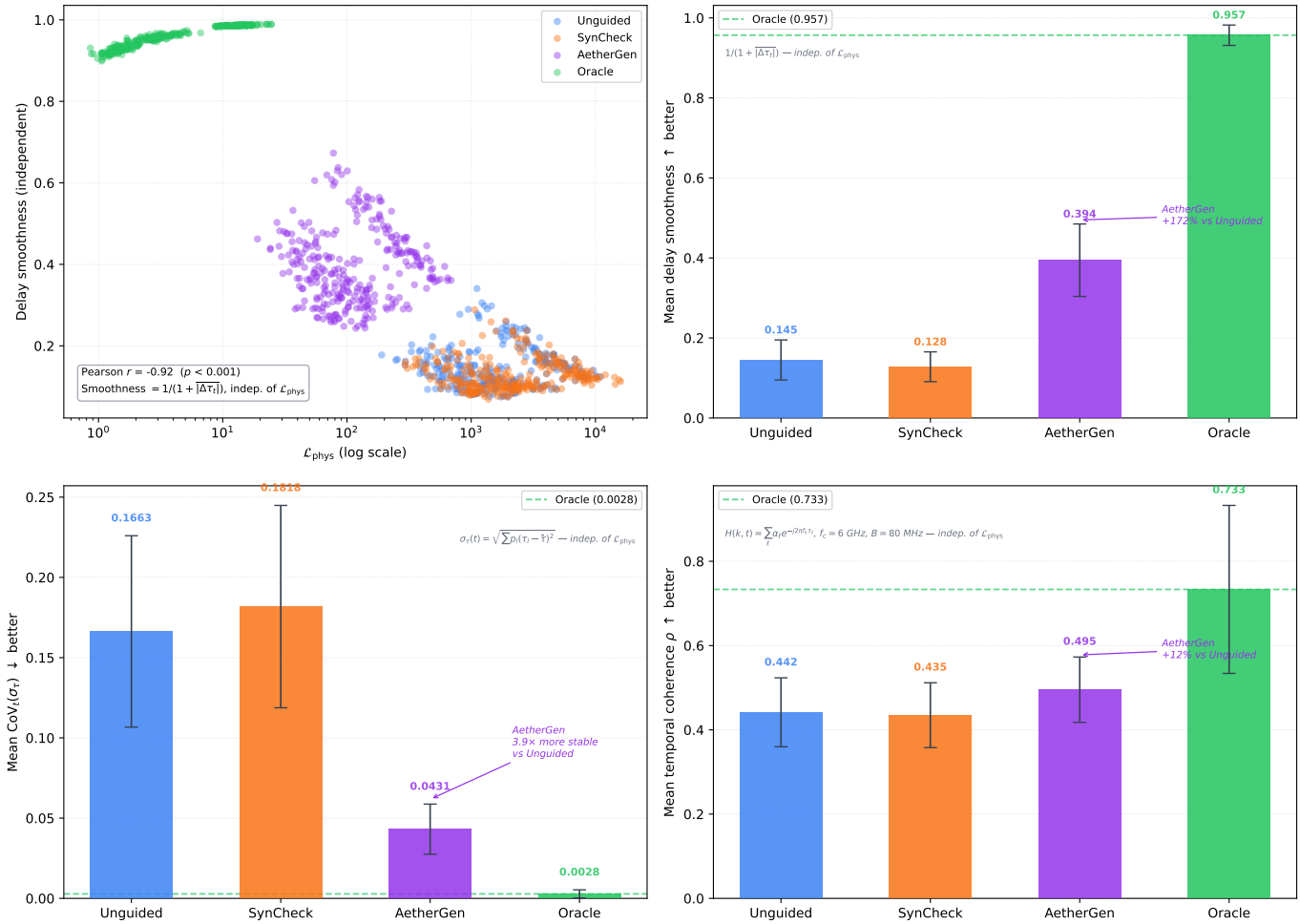

Fig. 5. **Cross-domain independent validation beyond $\mathcal{L}_{\text{phys}}$. (a)** Correlation between $\mathcal{L}_{\text{phys}}$ and an independent path-domain smoothness metric (Pearson r=−0.92, p<0.001), showing that physics-loss reductions correspond to genuine physical improvements. **(b)** Mean delay smoothness by condition; AetherGen improves substantially over Unguided and SynCheck, while post-hoc selection alone does not. **(c)** CIR-domain temporal stability $\text{CoV}_t(\sigma_\tau)$ (↓ better): AetherGen is 3.9× more stable than Unguided (0.043 vs. 0.166). **(d)** CSI-domain temporal coherence ρ (↑ better): AetherGen achieves +12% over Unguided (p<0.001, N=300). All four metrics are independent of $\mathcal{L}_{\text{phys}}$.

Fig. 7(b) increases the candidate pool size. ConflictRate remains high, showing that more samples do not mitigate the mismatch.

Data scalability analysis in Fig. 7 further supports the hypothesis that the conflict is an inherent formulation gap rather than a result of insufficient training.

H. Ablation Study

We perform two targeted ablations to validate the design of $\mathcal{L}_{\text{phys}}$.

a) A1: Loss component decomposition.: We re-evaluate ConflictRate under three weight configurations: (i) $\mathcal{L}_{\text{motion}}$ only ($\lambda_{\text{delay}} = \lambda_{\text{energy}} = \lambda_{\text{coupling}} = 0$), (ii) all components with equal weights (baseline), and (iii) all components *without* $\mathcal{L}_{\text{motion}}$. The threshold θ_p is recomputed as the median of $\mathcal{L}$ on Unguided samples for each configuration.

As shown in Fig. 8(A), configurations (i) and (ii) yield *identical* ConflictRate rankings: Unguided and SynCheck remain at ≈70%, while AetherGen and Oracle stay at 0%. This confirms that $\mathcal{L}_{\text{motion}}$ carries the complete discriminative signal. Under configuration (iii), Unguided and SynCheck drop to ≈52% while AetherGen rises to 100%—revealing that AetherGen specifically reduces $\mathcal{L}_{\text{motion}}$ and does not target the remaining components. This double-validation confirms temporal inconsistency as the singular root cause.

b) A4: Break-physics experiment.: To verify that each loss component measures a physically meaningful quantity rather than an artifact of the reference statistics, we deliberately construct physically invalid samples by applying three targeted perturbations to oracle samples: (1) *temporal disruption*: independent Gaussian noise added to each time frame of τ ($\sigma = 4\,\text{ns}$), destroying temporal coherence; (2) *coupling injection*: all coupling logits forced to 5.0 (sigmoid ≈ 0.99), making every path appear coupled regardless of geometry; (3) *delay spread*: a per-path constant offset sampled uniformly in $[0, 25\,\text{ns}]$ added

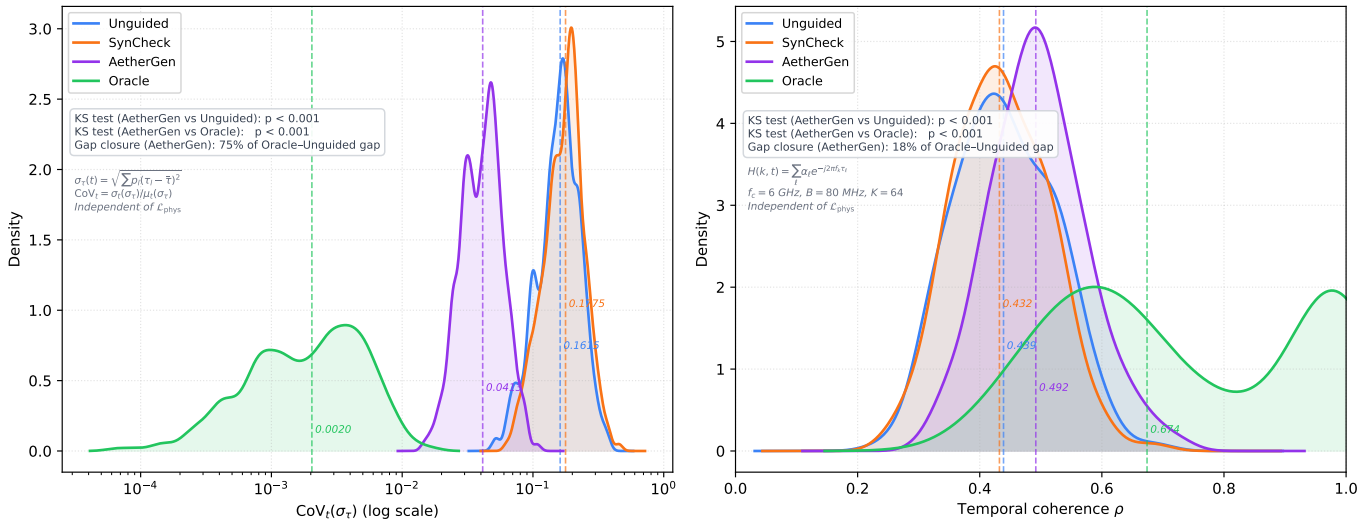

Fig. 6. Full sample distributions for the CIR- and CSI-domain metrics. Dashed lines mark per-condition medians. KS-test statistics confirm that AetherGen's distributions differ significantly from Unguided (p<0.001) in both domains. Gap closure rates quantify how much of the Oracle–Unguided gap is recovered: 75% in the CIR domain and 18% in the CSI domain, consistent with CSI being an indirect frequency-domain projection of path-level structure.

to all delays, shifting the delay distribution away from the reference.

Fig. 8(B) shows the fold-increase of each component relative to clean oracle samples (log scale). Temporal disruption increases $\mathcal{L}_{\text{motion}}$ by a factor exceeding $10^9\times$ while leaving $\mathcal{L}_{\text{coupling}}$ unchanged. Coupling injection increases $\mathcal{L}_{\text{coupling}}$ by $2.1\times$ while leaving all other terms unchanged. Delay spread increases $\mathcal{L}_{\text{delay}}$ by $73.7\times$ while leaving $\mathcal{L}_{\text{coupling}}$ and $\mathcal{L}_{\text{motion}}$ unchanged. This selective response demonstrates that each loss component is an independent measurement of a distinct physical property, directly rebutting the circular evaluation concern.

I. Summary of Evidence

Taken together, the results establish a coherent evidence chain:

- **Fig. 1:** Task-optimal samples are often physically invalid.
- **Fig. 2:** This mismatch is robust across environments and thresholds.
- **Fig. 3(a):** Post-hoc filtering fails to resolve the mismatch.
- **Fig. 3(b):** AetherGen resolves the conflict without trade-off.
- **Fig. 4:** Temporal inconsistency is identified as the primary failure mode, which AetherGen effectively mitigates by enforcing coherent path evolution.
- **Fig. 5:** Improvements are not metric-specific.
- **Fig. 7:** The issue is not due to model capacity or data scale.
- **Fig. 8:** $\mathcal{L}_{\text{motion}}$ is the singular root cause; each component measures an independent physical property (break-physics validation).

Final conclusion: The task–physics mismatch is a fundamental limitation of task-driven generation, and requires modifying the generative process itself.

VII. Mechanistic Analysis

In Sec. VI, we established that task-driven selection systematically favors samples that violate physical constraints. We now analyze the underlying mechanism behind this phenomenon.

Our analysis builds directly on the proposition introduced in Sec. VI-B: task objectives are invariant to structured transformations, while physical validity is not.

A. From Invariance to Structural Violation

Task models typically operate on aggregated or order-invariant representations, such as pooled features or summary statistics. As a result, they are insensitive to structured dependencies within the signal.

In contrast, physical validity in RF sensing depends critically on *structured relationships*, including temporal coherence and multipath consistency.

This mismatch creates an equivalence class:

$$\mathcal{T}(x) = \{x' : s_{\text{task}}(x') = s_{\text{task}}(x)\}, \tag{15}$$

within which the task model assigns identical scores, while $\mathcal{L}_{\text{phys}}(x')$ may vary significantly.

a) Implication.: Selection based on $s_{\text{task}}(x)$ is therefore unable to distinguish between physically valid and invalid configurations within $\mathcal{T}(x)$, leading to systematic selection bias.

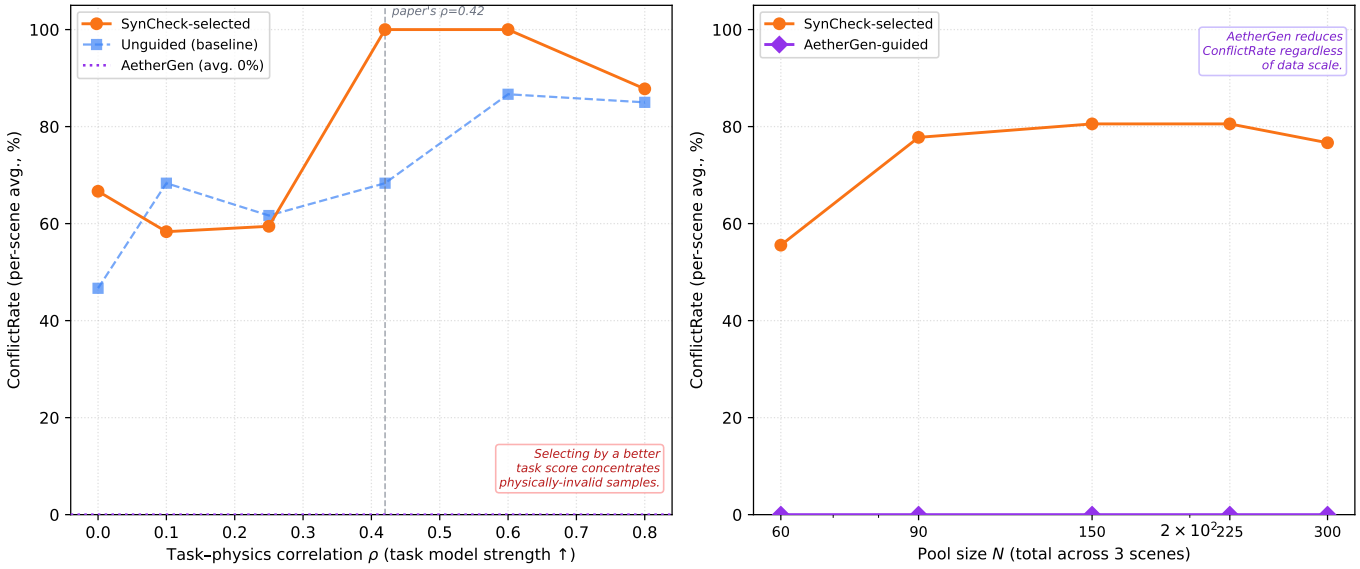

Fig. 7. **The mismatch is not due to model capacity or data scale.** (a) ConflictRate as a function of task model strength. Stronger task models increase, rather than reduce, the mismatch. (b) ConflictRate versus candidate pool size. Increasing sample size does not resolve the mismatch. These results indicate that the task–physics conflict is a formulation-level issue rather than a limitation of model capacity or data availability.

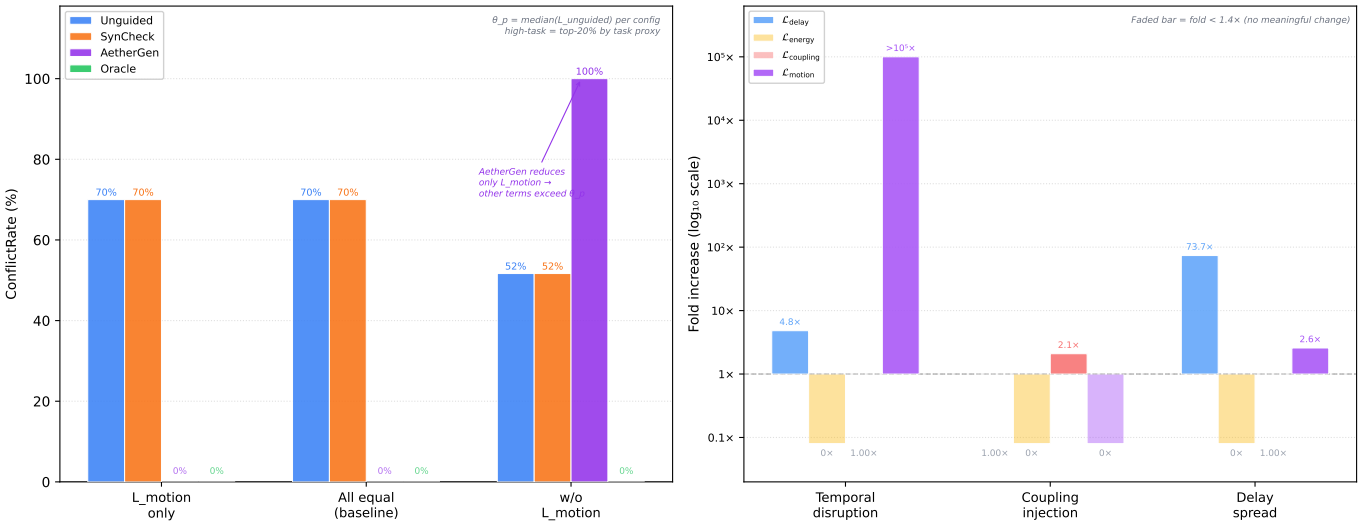

Fig. 8. Ablation study. **(A)** Loss decomposition: ConflictRate under three weight configurations. $\mathcal{L}_{\text{motion}}$ alone recovers the full discriminative signal; removing it collapses CR for Unguided/SynCheck while raising AetherGen's CR to 100%, confirming AetherGen targets only the temporal failure mode. **(B)** Break-physics: fold-increase of each loss component under targeted perturbations (log scale). Each perturbation selectively elevates its designated component, demonstrating that $\mathcal{L}_{\text{phys}}$ captures real physical structure independent of the training objective. Hatched bars indicate near-zero baselines where absolute fold is unreliable.

B. Dominance of Temporal Inconsistency

To identify which physical constraints are most affected, we decompose the physics loss:

$$\mathcal{L}_{\text{phys}} = \mathcal{L}_{\text{delay}} + \mathcal{L}_{\text{energy}} + \mathcal{L}_{\text{coupling}} + \mathcal{L}_{\text{motion}}. \tag{16}$$

Fig. 4(a) shows that $\mathcal{L}_{\text{motion}}$ dominates the total loss, accounting for more than 99% of physical violations. This suggests that while the generative model captures static feature distributions, it fails to learn the underlying causal dynamics of path evolution.

Fig. 4(b) illustrates the distribution gap: unguided samples exhibit a high-variance tail in temporal evolution, which AetherGen effectively shifts back toward the oracle distribution.

This indicates that *temporal inconsistency* is the primary failure mode of task-driven generation in our setting.

C. Why Temporal Structure Is Ignored

Temporal coherence is a structured dependency that requires consistency across time frames. However, task models operating on aggregated representations do not enforce such constraints.

As a result, two samples with identical task scores may exhibit vastly different temporal behaviors:

- one with smooth, physically plausible evolution,
- another with abrupt, inconsistent variations.

Because the task objective does not penalize these differences, selection favors samples that are optimal in task space but inconsistent in temporal structure.

D. Resolution via Physics-Guided Generation

AetherGen addresses this issue by explicitly incorporating structured constraints during generation.

By injecting gradients of $\mathcal{L}_{\text{phys}}$ into the sampling process, AetherGen enforces consistency within each equivalence class $\mathcal{T}(x)$, effectively selecting physically valid representatives.

Empirically, this results in:

- reduced temporal variance,
- lower $\mathcal{L}_{\text{motion}}$,
- consistent improvements in overall physical validity.

E. Summary

The task–physics conflict arises from a fundamental mismatch:

$$\text{task invariance} \Rightarrow \text{structural ambiguity}. \tag{17}$$

Without explicit constraints, this ambiguity leads to the selection of physically invalid samples.

Physics-guided generation resolves this issue by enforcing structured consistency during sampling, aligning task utility with physical validity.

VIII. Discussion

A. Beyond RF Sensing: A General Limitation of Task-Driven Generation

While this work focuses on RF signal generation, the identified task–physics mismatch reflects a broader limitation of task-driven generative modeling.

As shown in Sec. VII, the root cause lies in the invariance of task objectives to structured dependencies. When task models operate on aggregated or order-invariant features, they fail to penalize incoherent structure in the generated outputs. This leads to a systematic gap between task-optimality and structured validity.

Importantly, this phenomenon is not specific to RF sensing. Similar conditions arise in many domains: temporal consistency in video generation, spatial coherence in 3D scene synthesis, and relational consistency in structured data generation.

Our results suggest that, in such settings, optimizing task objectives alone is insufficient to guarantee meaningful or physically valid outputs.

B. Implications for Generative Model Design

The findings of this work point to a key design principle:

> *Structured dependencies must be explicitly enforced, rather than implicitly assumed.*

In conventional pipelines, physical or structural validity is often treated as a secondary concern, addressed through post-hoc filtering or selection. However, our experiments demonstrate that such approaches cannot resolve structural inconsistencies that originate during generation.

Our findings advocate for integrating constraints directly into the generative process, rather than relying on post-hoc interventions. AetherGen provides one such approach by injecting physics-based gradients during sampling, effectively shaping the generative trajectory.

More broadly, this suggests a paradigm shift: *generation should be guided by both data and structure, not data alone.*

C. Relationship to Prior Approaches

Existing methods for improving generative quality typically fall into two categories: (1) enhancing model capacity (e.g., larger networks or more data), and (2) post-hoc selection or filtering.

Our results show that neither approach addresses the task–physics mismatch. As demonstrated in Sec. VI, increasing model strength or sample pool size does not reduce the ConflictRate, and filtering can even exacerbate physical violations.

This highlights a fundamental distinction: improving *representation power* does not guarantee *structural correctness.*

In contrast, AetherGen directly targets the source of inconsistency by enforcing structured constraints during generation.

D. Limitations

Despite its effectiveness, AetherGen has several limitations.

First, our experiments are conducted on synthetic data with controlled scene configurations. While this enables precise analysis, real-world RF environments may introduce additional complexities beyond the scope of this study.

Second, the current formulation focuses on temporal consistency, which is the dominant failure mode in our setting. Other structural constraints, such as path ordering or geometric consistency, are not explicitly enforced.

Third, the strength of guidance (λ) introduces a trade-off between constraint enforcement and sample diversity. While we observe stable improvements in our experiments, adaptive guidance strategies may further improve robustness.

E. Future Directions

Several directions emerge from this work.

First, extending physics-guided generation to real-world datasets (e.g., measured CSI or radar signals) is an important next step.

Second, incorporating richer structural constraints—such as geometry-aware propagation or learned material properties—may further improve physical fidelity.

Third, exploring alternative forms of constraint injection, including learned or hybrid approaches, could generalize the framework beyond hand-designed physics losses.

Finally, we believe that the task–structure mismatch identified here is relevant to a broader class of generative problems. Investigating similar phenomena in domains such as video, robotics, and scientific simulation is a promising avenue for future research.

F. Closing Perspective

This work highlights a fundamental gap in generative modeling:

$$\text{Task-optimality} \not\Rightarrow \text{Structured validity.} \tag{18}$$

Bridging this gap requires moving beyond purely task-driven objectives and explicitly incorporating domain structure into the generation process.

We hope that this perspective encourages a shift toward structure-aware generative modeling, where validity is enforced by design rather than assumed as a byproduct of task optimization.

IX. Conclusion

In this work, we investigated the problem of generating physically consistent RF path signals and identified a fundamental limitation of task-driven generative modeling: optimizing for task utility does not guarantee physical validity.

Through controlled experiments, we revealed a systematic *task–physics mismatch*, where samples with high task confidence often exhibit severe violations of physical consistency. We further showed that this mismatch is robust across environments and thresholds, and cannot be resolved by post-hoc selection or increased model capacity.

To address this issue, we proposed **AetherGen**, a physics-guided diffusion framework that enforces physical constraints during the generation process. Unlike filtering-based approaches, AetherGen directly modifies the generative trajectory, producing samples that are both task-effective and physically consistent.

A mechanistic analysis revealed that temporal inconsistency is the dominant failure mode of unconstrained generation, and that the root cause lies in the invariance of task objectives to structured dependencies. Physics guidance resolves this issue by injecting structured constraints into the sampling process, enforcing coherent evolution over time.

More broadly, our study reveals a fundamental decoupling between task-optimality and structural validity in generative RF sensing.

Bridging this gap requires explicitly incorporating domain structure into generative models, rather than relying solely on task-driven optimization. We believe this perspective extends beyond RF sensing to a wide range of generative tasks where structural consistency is critical.

Future work will focus on extending physics-guided generation to real-world data, incorporating richer structural constraints, and exploring more general forms of constraint-aware generative modeling.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[3] X. Chen and X. Zhang, "RF genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models," in *AProceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, 2023, pp. 28–42.

[4] C. Gong, B. Liang, W. Gao, and C. Xu, "Data can speak for itself: Quality-guided utilization of wireless synthetic data," in *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, 2025, pp. 209–222.

[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[6] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2020.

[7] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[8] A. A. M. Saleh and R. A. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 2, pp. 128–137, 1987.

[9] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Report TR 38.901 V17.0.0, Mar. 2022, release 17.

[10] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. De Doncker, "The COST 2100 MIMO channel model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, 2012.

[11] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," *arXiv preprint arXiv:1902.06435*, 2019.

[12] J. Hoydis, F. Aït Aoudia, S. Cammerer, M. Nimier-David, N. Binder, G. Marcus, and A. Keller, "Sionna RT: Differentiable ray tracing for radio propagation modeling," in *2023 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2023, pp. 317–321.

[13] L. U. Khan, Z. Han, W. Saad, E. Hossain, M. Guizani, and C. S. Hong, "Digital twin of wireless systems: Overview, taxonomy, challenges, and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2230–2254, 2022.

[14] IEEE, "IEEE standard for information technology–telecommunications and information exchange between systems local and metropolitan area networks–specific requirements–part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications–amendment 1: Enhancements for high-efficiency WLAN," IEEE Std 802.11ax-2021, 2021.