DynHD: Hallucination Detection for Diffusion Large Language Models via Denoising Dynamics Deviation Learning

Anonymous ACL submission

Abstract

Diffusion large language models (D-LLMs) have emerged as a promising alternative to autoregressive models due to their iterative refinement capabilities. However, hallucinations remain a critical issue that hinders their reliability. To detect hallucination responses from model outputs, token-level uncertainty (e.g., entropy) has been widely used as an effective signal to indicate potential factual errors. Nevertheless, the fixed-length generation paradigm of D-LLMs implies that tokens contribute unevenly to hallucination detection, with only a small subset providing meaningful signals. Moreover, the evolution trend of uncertainty throughout the diffusion process can also provide important signals, highlighting the necessity of modeling its denoising dynamics for hallucination detection. In this paper, we propose DynHD that bridge these gaps from both *spatial* (token sequence) and *temporal* (denoising dynamics) perspectives. To address the information density imbalance across tokens, we propose a semantic-aware evidence construction module that extracts hallucination-indicative signals by filtering out non-informative tokens and emphasizing semantically meaningful ones. To model denoising dynamics for hallucination detection, we introduce a reference evidence generator that learns the expected evolution trajectory of uncertainty evidence, along with a deviation-based hallucination detector that makes predictions by measuring the discrepancy between the observed and reference trajectories. Extensive experiments demonstrate that DynHD consistently outperforms state-of-the-art baselines while achieving higher efficiency across multiple benchmarks and backbone models.

1 Introduction

Diffusion large language models (D-LLMs) (Nie et al., 2025; Ye et al., 2025) have recently received extensive attention as a significant generative paradigm alternative to autoregressive large language models (AR-LLMs) (Achiam et al., 2023; Vaswani et al., 2017). Unlike AR-LLMs using token-by-token decoding, D-LLMs generate fix-length sequences through multi-step iterative denoising, performing discrete remasking at each step: the model predicts a denoised intermediate sequence while selectively remasking a portion of positions for further optimization (Austin et al., 2021; Sahoo et al., 2024). This paradigm offers advantages in global consistency and the ability to gradually refine the generated contents (Li et al., 2022; Huang et al., 2024).

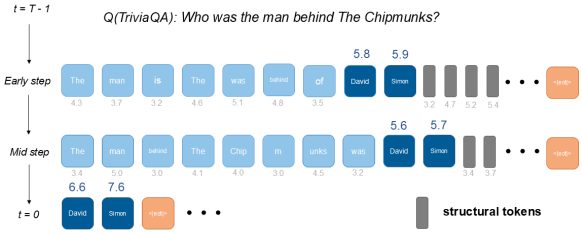


Figure 1: Visualization of spatial uncertainty distribution during decoding. Tokens exhibiting the highest entropy spikes serve as the primary indicators of factual instability. On the contrary, intermediate and structural tokens provide limited cues for hallucination detection.

While D-LLMs offer several advantages, they still suffer from hallucinations, i.e., generating factually incorrect or unsupported content, which limits their reliability in real-world applications (Ji et al., 2025; Li et al., 2025). To mitigate this issue, hallucination detection has become a critical task for ensuring the factual reliability of D-LLMs. Nevertheless, although extensive efforts have been devoted to the AR-LLMs setting, these hallucination detection approaches do not readily generalize to D-LLMs due to their fundamentally different architectural and generative paradigms (Yehuda et al., 2024). In particular, the key evidence identifying hallucinations in D-LLMs is typically distributed across the dynamic evolution of the en-

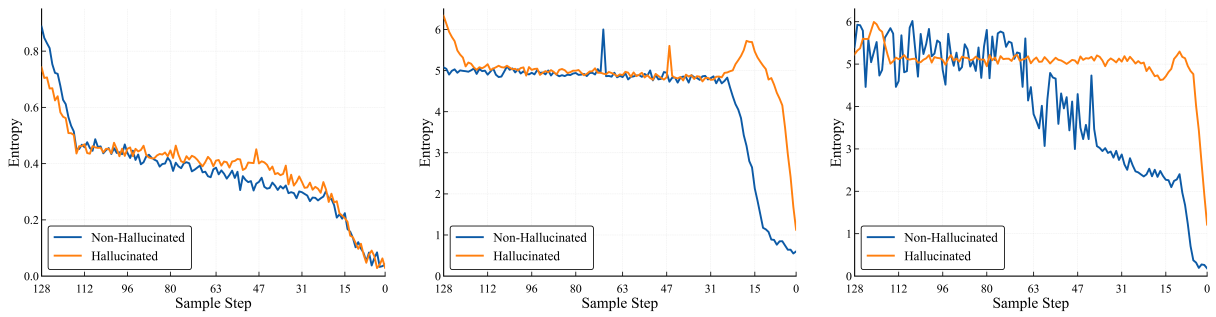


Figure 2: Entropy dynamics under spatial filtering. (a) Mean trajectories fail to distinguish between samples due to information density imbalance. (b) & (c) Filtered trajectories (top-K) reveal distinct convergence: factual samples decay monotonically, while hallucinated ones exhibit late-stage stagnation or rebound. Differences in decay rates between HotpotQA (b) and TriviaQA (c) reflect varying task-specific denoising dynamics.

tire denoising trajectory, rather than being concentrated in the output confidence of the final step. Consequently, hallucination detectors designed for AR-LLMs, such as those relying on final token probabilities (Kadavath et al., 2022; Kossen et al., 2024), single-step uncertainty adn entropy (Du et al., 2024), or multi-sample consistency (Manakul et al., 2023), often ignore the process-level signals inherent in the denoising trajectory. Moreover, these approaches, which are already computationally expensive for AR-LLMs, become even more time-consuming for D-LLMs, as additional sampling or verification requires repeatedly executing the denoising process. (Chern et al., 2023).

As a pioneering work on hallucination detection for D-LLMs, TraceDet (Chang et al., 2026) adopts a trajectory-based perspective by treating the denoising process as observable action traces (Black et al., 2023). It extracts hallucination-relevant sub-trajectories for discrimination and emphasizes the importance of "process signals" over "final-state signals" for hallucination detection in D-LLMs. However, through our in-depth analysis of the hallucination-indicated signals in D-LLMs, TraceDet still exhibits two orthogonal weaknesses: From a spatial perspective, it ignores the **information density imbalance** within generated sequences, failing to leverage the few tokens where hallucination signals are concentrated. From a temporal perspective, it overlooks the underlying **denoising dynamics** of entropy-based hallucination evidence, hindering the modeling of uncertainty evolution in the diffusion trajectory.

Firstly, regarding the spatial dimension, D-LLMs naturally generate answers within fixed-length sequences (Li et al., 2022; Sahoo et al., 2024). In these sequences, the roles of individual tokens are highly heterogeneous, leading to an **information density imbalance** (Gong et al., 2023; Gu et al., 2018), as illustrated in Figure 1. A few tokens are vital for the final answer, providing key signals for hallucination detection. In contrast, the majority of tokens are merely intermediate guesses that are later remasked, or structural padding to introduced to maintain the fixed-length sequence, which contain little hallucination-related information. In the context of such imbalance, TraceDet (Chang et al., 2026) aggregates uncertainty from all tokens as detection evidence without explicitly accounting for their semantic importance, which may drown hallucination signals with low-informative tokens (Zhang et al., 2023). As a result, the entropy profiles of hallucinated and factual samples become indistinguishable to each other (see Figure 2a), leading to sub-optimal hallucination detection performance.

Even if hallucination-related evidence can be extracted from token sequences, how to leverage the evidence trajectory over multiple diffusion steps remains another challenge. Through visualization of evidence trajectories, we empirically find that **denoising dynamics**, i.e., the trend of how evidence evolves throughout the diffusion process, are critical for hallucination detection. As shown in Figure 2b and Figure 2c (corresponding examples are in Appendix D.1), factual and hallucinated samples often show different trends during the denoising process, providing strong discriminative signals. More importantly, the range, trend, and shape of entropy curves often vary significantly across different scenarios and questions (see the difference between Figures 2b and 2c), highlighting the ne-

cessity of modeling complex denoising dynamics for robust hallucination detection. Nevertheless, TraceDet selects only a few key steps for hallucination prediction, which may neglect the continuous and holistic dynamics of the diffusion process.

To overcome the above limitations, we propose **DynHD**, a denoising **Dyn**amics deviation learning-based approach for **H**allucination **D**etection of D-LLMs. To address the *information density imbalance*, our method not only introduces a semantic-aware token filtering module to remove structural and non-informative tokens for reliable evidence extraction, but also leverages multivariate statistical modeling of entropy to construct evidence with strong hallucination indications. Moreover, we model denoising dynamics for robust hallucination detection with a denoising dynamical deviation learning module, where a reference evidence dynamics generator is established to model the normal evolution patterns of evidence, and a deviation-based detector identifies hallucinations by measuring the difference between observed evidence dynamics and the reference dynamics. Extensive experiments on TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), and CommonsenseQA (Talmor et al., 2019) with Dream-7B (Ye et al., 2025) and LLaDA-8B (Nie et al., 2025) demonstrate that DynHD achieves state-of-the-art performance. It outperforms TraceDet by an average AUROC margin of 12.2% without requiring external retrieval or computationally expensive repeated sampling.

2 Preliminaries

Diffusion Denoising Mechanism. Diffusion large language models (D-LLMs) generate sequences via an iterative stochastic denoising process. Given a prompt $\mathbf{q}$, the model maintains a discrete state $\mathbf{x}^{(t)} = (x_t^{(1)}, \ldots, x_t^{(l)}) \in \mathcal{V}^l$ across the trajectory $t \in \{T, \ldots, 0\}$, where $\mathcal{V}, l, T$ denote the vocabulary, fixed sequence length, and total denoising steps, respectively. Starting from a highly masked initialization $\mathbf{x}^{(T)}$, the model iteratively fills masked positions to reach the final output $\mathbf{x}^{(0)}$.

At each step t, the model predicts a categorical distribution $\pi_i^{(t)} \in \mathbb{R}^{|\mathcal{V}|}$ for each position i via a softmax over the generated logits. We quantify the position-wise prediction uncertainty using the token entropy:

$$s_{t,i} = -\sum_{v \in \mathcal{V}} \pi_i^{(t)}(v) \log \pi_i^{(t)}(v). \tag{1}$$

Let $\mathcal{S}_t = \{s_{t,i}\}_{i=1}^{l}$ represent the sequence of token entropies at step t. The complete raw uncertainty trajectory spanning all descending denoising steps is thus denoted as $\mathcal{T} = (\mathcal{S}_T, \mathcal{S}_{T-1}, \ldots, \mathcal{S}_0)$. Notably, due to the fixed length l, sequences that terminate early are padded with non-semantic structural tokens ($\langle|endoftext|\rangle$), which introduces significant noise into $\mathcal{T}$.

Definition of Hallucination Detection Problem. We formulate hallucination detection as a binary classification task for a question-response pair $(\mathbf{q}, \mathbf{r})$, where $\mathbf{r} = \mathbf{x}^{(0)}$. Given a training set $\mathcal{D} = \{(\mathbf{q}_n, \mathbf{r}_n, y_n)\}_{n=1}^{N}$ with labels $y_n \in \{0, 1\}$ assigned by an automatic judge (Zheng et al., 2023), where $y = 1$ indicates a hallucinated response and $y = 0$ denotes a factually correct one. Our goal is to learn an optimal detector h_ϕ (e.g., a neural network parameterized by ϕ) that minimizes the empirical risk:

$$\begin{aligned} \min_{\phi} \quad & \mathbb{E}_{(\mathbf{q},\mathbf{r},y)\sim\mathcal{D}} \left[\mathcal{L}(y, h_\phi(\mathbf{q}, \mathbf{r}, \mathcal{T}))\right], \\ \text{s.t.} \quad & \mathbf{x}^{(t-1)} \sim p_\theta(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}, \mathbf{q}), \end{aligned} \tag{2}$$

where $\mathcal{L}$ is a classification loss function, and the detector makes its decision based on the question, the final response, and the collected uncertainty trajectory $\mathcal{T}$.

The related work is summarized in Appendix A.

3 Methodology

In this section, we introduce DynHD that detects hallucinations by learning deviations in denoising dynamics. As illustrated in Figure 3, DynHD starts with a *semantic-aware evidence construction* module (Sec. 3.1). To handle the information density imbalance issue, it first filters out non-informative structural token, and then constructs the hallucination evidence by statistically aggregating the uncertainty of the remaining semantic tokens. After that, a *dynamical deviation learning* module (Sec. 3.2) identifies the hallucination answers by modeling denoising dynamics of D-LLMs. Concretely, we model the continuous denoising process by building a question-conditioned reference path to serve as a baseline, and then compare the actual evidence trajectory against this established reference for hallucination prediction.

3.1 Semantic-aware Evidence Construction

To identify hallucinations generated by D-LLMs, the first step is to construct hallucination-indicative

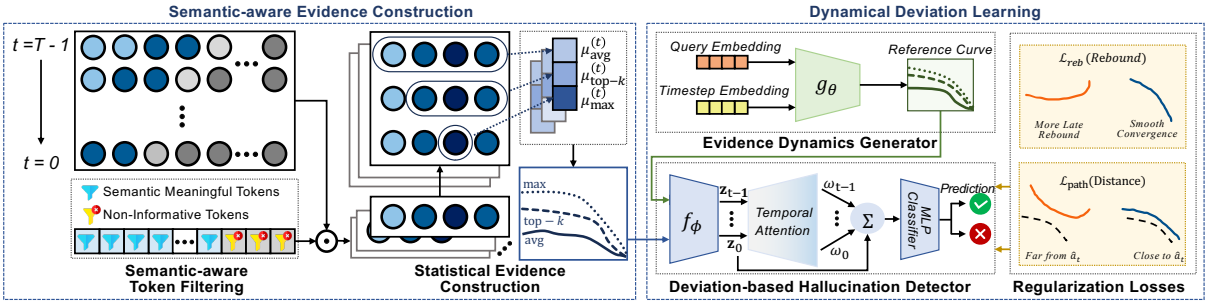


Figure 3: Framework overview of DynHD, which constructs a semantic-aware evidence trajectory from uncertainty, and detects hallucination via learning the deviation between the reference and observed trajectories.

evidence from the denoising process. While token-level uncertainty (e.g., entropy) can serve as a useful indicator, the fixed-length generation paradigm of D-LLMs introduces an information density imbalance, i.e., not all tokens in a generated sequence provide meaningful signals for hallucination detection. To address this issue, we adopt a two-stage strategy for evidence construction. First, we filter out non-informative structural tokens using a rule-based filtering mechanism. Then, we construct statistical evidence by aggregating the entropy of the remaining tokens with the consideration of their their semantic importance in the sequence.

Semantic-aware Token Filtering. Due to the inherent fixed-length generation paradigm of D-LLMs, a large number of tokens within the generated sequence are often non-informative, especially when the final answer is short. These structural and meaningless tokens often have little relevance to the final answer, and therefore provide little useful signal for hallucination detection.

To extract semantically relevant uncertainty for hallucination detection, we first employ a filtering mechanism to directly exclude these non-informative tokens from the evidence calculation. We define an filtering set $\mathcal{I}_{\text{ignore}}$ to systematically filter out these tokens across five linguistic dimensions: ❶ Control & Boundary Markers (e.g., <|endoftext|>, [PAD]); ❷ Lexical Noise (e.g., punctuation and whitespace) causing vacuous entropy spikes; ❸ Task Boilerplate (e.g., "Answer:") that deflates mean entropy with deterministic patterns; ❹ Stopwords yielding minimal information; and ❺ Subword Fragments constrained by prefixes. By discarding this uninformative majority, we leave only the informative tokens to participate in the evidence computation. At each denoising step t, for token $x_i^{(t)}$ at position $i \in \{1, \ldots, l\}$ (where l is the sequence length), we extract the subset of valid semantic positions $\mathcal{K}_t$:

$$\mathcal{K}_t = \{i \mid 1 \leq i \leq l,\, x_i^{(t)} \notin \mathcal{I}_{\text{ignore}}\}. \tag{3}$$

Confining the evidence construction to $\mathcal{K}_t$ ensures that our extracted signals are grounded purely in semantic signals rather than formatting artifacts.

Statistical Evidence Construction. While the filtering mechanism successfully removes non-informative tokens, the remaining semantic tokens in $\mathcal{K}_t$ still contribute differentially to hallucinations (e.g., factual nouns versus prepositions or pronouns). Consequently, a simple average of these tokens may smooth out localized signals and produces weak evidence. To better capture informative patterns, we adopt a statistics-based method to condense the purified uncertainty field $\{s_{t,i}\}_{i \in \mathcal{K}_t}$ at each step t into a statistical evidence vector $\mathbf{a}_t \in \mathbb{R}^3$. This captures the mean, peak, and top-k characteristics of the entropy distribution:

$$\mathbf{a}_t = \left[\mu_{\text{avg}}^{(t)},\, s_{\text{max}}^{(t)},\, \mu_{\text{top-}k}^{(t)}\right]. \tag{4}$$

These three statistical moments capture the semantic uncertainty at different granularities: ❶ **Global Uncertainty** ($\mu_{\text{avg}}^{(t)}$) tracks the mean uncertainty across $\mathcal{K}_t$, indicating macro-level convergence to a coherent state; ❷ **Peak Uncertainty** ($s_{\text{max}}^{(t)}$) captures the maximum entropy spike, revealing micro-level factual errors (e.g., incorrect entities) that typically masked by global averaging; ❸ **Regional Uncertainty** ($\mu_{\text{top-}k}^{(t)}$) averages the top-k entropy (or all, if $|\mathcal{K}_t| < k$), capturing the phrase scale uncertainties of semantic spans. Together, these statistics allow the detector to capture errors at different scales and remain robust to varying sequence lengths. This approach ensures that isolated factual mistakes are not hidden within fluent and confident contexts.

Finally, stacking these step-wise statistics transforms the raw trajectory $\mathcal{T}$ into a compact evidence trajectory $\mathcal{E} = (\mathbf{a}_T, \mathbf{a}_{T-1}, \ldots, \mathbf{a}_0)$, mapping the semantic uncertainty resolution from initial noise ($t = T$) to the final generation ($t = 0$).

3.2 Dynamical Deviation Learning

Once high-quality evidence is constructed, the next question is how to leverage it for hallucination detection. While the above module focuses on the spatial dimension (i.e., token-level filtering) of evidence construction, the temporal dimension is also crucial, namely how to model the denoising dynamics of the evidence to identify hallucinations. Although the evolution of evidence can reveal useful patterns, the range and trend of entropy may vary significantly depending on the specific scenarios and questions, making it difficult to directly use the evidence for hallucination prediction. To mine discriminative signals from denoising dynamics of evidence, we propose a dynamical deviation learning module consisting of two components. First, we build an *evidence dynamics generator* to predict an expected reference trajectory conditioned on the input question, which captures the specific evolution patterns of evidence under different Q&A scenarios. Meanwhile, a *deviation-based hallucination detector* makes the final prediction by adaptively measuring the discrepancy between the actual evidence trajectory $\mathcal{E}$ and the predicted reference, supporting reliable hallucination detection.

Reference Evidence Dynamics Generator. Since denoising dynamics vary significantly across different queries and scenarios, raw evidence is not a reliable direct indicator for hallucination detection. Rather than predicting hallucinations directly from evidence, we design a generator g_θ to model the expected evidence trajectory. Conditioned on the input query embedding $\mathbf{q}$ and the timestep embedding $\mathbf{e}_t$, g_θ learns the relationship between the task content and the uncertainty dynamics. This produces a reference trajectory that represents normal, factual behavior to assist the detection process.

For each step t, the generator predicts a reference evidence vector $\hat{\mathbf{a}}_t = g_\theta(\mathbf{q}, \mathbf{e}_t)$. We optimize the generator using only factual samples ($y = 0$) via:

$$\mathcal{L}_{\text{ref}} = \mathbb{E}_{y=0}\left[\frac{1}{T}\sum_{t=0}^{T}\text{SmoothL1}(\mathbf{a}_t, \hat{\mathbf{a}}_t)\right]. \quad (5)$$

This approach allows the framework to model task-specific dynamics of factual sequences. In practice, we first pre-train g_θ using $\mathcal{L}_{\text{ref}}$ to establish a stable reference space for the subsequent detection stage.

Deviation-based Hallucination Detector. Based on the reference evidence trajectory, we use the deviation between the actual and reference curves as the direct criterion for hallucination detection. The detector can infer hallucinations by adaptively evaluating differences in curve trends. Specifically, we first construct a composite feature vector $\mathbf{x}_t = [\mathbf{a}_t;\ \hat{\mathbf{a}}_t;\ \Delta\mathbf{a}_t]$ at each step t to model the dynamics. Here, $\mathbf{a}_t$ and $\hat{\mathbf{a}}_t$ represent the observed and reference states, while $\Delta\mathbf{a}_t = \mathbf{a}_{t-1} - \mathbf{a}_t$ captures the step-wise velocity toward the subsequent state. A shared MLP, f_ϕ, then projects these vectors into hidden representations $\mathbf{z}_t$.

Since different steps contribute differently to detection, we employ a learnable weight ω_t to model their importance. A scoring network evaluates each time step by computing unnormalized scores

$$u_t = \mathbf{w}^\top \tanh(\mathbf{U}\mathbf{z}_t + \mathbf{b}_u), \quad (6)$$

where $\mathbf{w}$, $\mathbf{U}$, and $\mathbf{b}_u$ are learnable parameters. Then, these scores are normalized via a softmax function to produce the weight distribution $\omega_t = \exp(u_t)/\sum_{j=0}^{T}\exp(u_j)$. Finally, the hallucination prediction $\tilde{y}$ is computed by $\tilde{y} = \text{MLP}\left(\sum_{t=0}^{T}\omega_t\mathbf{z}_t\right)$, where temporal phases with prominent deviations can be prioritized.

Regularization Losses. To better identify hallucination characteristics, we introduce two dynamical regularizers weighted by ω_t that focus on late-stage stagnation and uncertainty rebound. Concretely, the path-deviation score s_{path} quantifies cumulative divergence:

$$s_{\text{path}} = \sum_{t=0}^{T}\omega_t\sum_{d=1}^{D}|a_{t,d} - \hat{a}_{t,d}|. \quad (7)$$

The rebound score s_{reb} captures uncertainty increases along the descending steps ($T \to 0$):

$$s_{\text{reb}} = \sum_{t=1}^{T}\omega_t\sum_{d=1}^{D}\big(\max(0, a_{t-1,d} - a_{t,d})\big)^2. \quad (8)$$

We define an adaptive margin $\hat{m}^{(\tau)}$ to represent the boundary of factual behaviors, updated via EMA (He et al., 2020) using factual samples:

$$\begin{cases} m_{\text{batch}} = \text{Quantile}(\{s_i \mid y_i = 0\}, q) \\ \hat{m}^{(\tau)} = (1-\beta)\hat{m}^{(\tau-1)} + \beta m_{\text{batch}}. \end{cases} \quad (9)$$

Table 1: AUROC(%) comparison of hallucination detection methods on two D-LLMs across three QA datasets. The highest score is **bolded** and the second highest is underlined.

Model Method	TriviaQA		HotpotQA		CSQA		Avg
	128	64	128	64	128	64	
LLaDA-8B-Instruct							
Output-based Methods							
Perplexity	50.4	47.6	49.3	51.2	65.6	65.0	54.9
LN-Entropy	54.6	53.5	54.8	54.7	64.6	64.4	57.8
Semantic Entropy	68.9	67.3	57.6	53.8	44.1	43.9	55.9
Lexical Similarity	62.5	59.0	64.2	57.1	57.3	60.7	60.1
Latent-based Methods							
EigenScore	69.2	66.9	64.7	59.2	58.5	60.6	63.2
CCS	57.1	54.2	57.6	55.8	50.5	58.5	55.6
TSV	60.2	61.1	65.0	59.4	52.9	55.2	59.0
Trajectory-based Methods							
TraceDet	73.9	74.1	66.1	63.7	77.2	77.1	72.0
DynHD	**86.7**	**86.1**	**84.2**	**85.3**	**81.6**	**81.3**	**84.2**
Dream-7B-Instruct							
Output-based Methods							
Semantic Entropy	73.7	72.5	62.7	67.7	51.4	48.6	62.8
Lexical Similarity	58.3	64.0	59.7	62.7	77.3	76.9	66.5
Latent-based Methods							
EigenScore	66.0	69.1	62.5	67.0	76.9	77.5	69.8
CCS	56.9	50.3	51.7	58.2	54.2	53.2	54.1
TSV	75.6	74.7	58.7	63.0	62.3	56.8	65.2
Trajectory-based Methods							
TraceDet	78.1	**86.7**	75.1	76.0	**84.7**	84.1	80.8
DynHD	**87.3**	84.4	**80.1**	**85.6**	83.5	**84.6**	**84.3**

The regularization losses are formulated as:

$$\mathcal{L}_{\text{path}} = \mathcal{L}_{\text{hinge}}(s_{\text{path}}; \hat{m}^{(\tau)}_{\text{path}}), \tag{10}$$

$$\mathcal{L}_{\text{reb}} = \mathcal{L}_{\text{hinge}}(s_{\text{reb}}; \hat{m}^{(\tau)}_{\text{reb}}), \tag{11}$$

where $\mathcal{L}_{\text{hinge}}(s; \hat{m}) = (1-y) \cdot s + y \cdot \max(0, \hat{m} - s)$ (Cortes and Vapnik, 1995). This forces the model to minimize deviations for factual samples while pushing hallucinations beyond the boundary. The total training objective combines the classification loss $\mathcal{L}_{\text{cls}}$ with these regularizers:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{path}} + \lambda_2 \mathcal{L}_{\text{reb}}, \tag{12}$$

where λ_1 and λ_2 are balance hyperparameters. Detailed pseudocode and complexity analysis of DynHD are provided in Appendix B.

4 Experiments

4.1 Experimental Setup

Datasets and Models. To ensure a fair and rigorous comparison, we strictly adhere to the evaluation protocol established by TraceDet (Chang et al., 2026). Specifically, our experiments are conducted across three widely adopted fact-based QA benchmarks: TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), and CommonsenseQA (Talmor et al., 2019), which cover open-domain fact-based questions, multi-hop reasoning, and commonsense reasoning. We adopt two representative models, LLaDA-8B-Instruct (Nie et al., 2025) and Dream-7B-Instruct (Ye et al., 2025), as backbone D-LLMs.

Baselines. We compare our method against the same set of baselines as TraceDet, covering both output-based and latent-based hallucination detection approaches. Output-based baselines include Perplexity (Ren et al., 2023), Length-Normalized Entropy (LN-Entropy) (Malinin and Gales, 2020), Semantic Entropy (Kuhn et al., 2023), and Lexical Similarity (Lin et al., 2024). Latent-based baselines include EigenScore (Chen et al., 2024), Contrast-Consistent Search (CCS) (Burns et al., 2023), and Truthfulness Separator Vector (TSV) (Park et al., 2025). We also compare our framework against TraceDet (Chang et al., 2026), the state-of-the-art method for D-LLM hallucination detection.

Evaluation Protocol. We evaluate hallucination detection performance using AUROC, following prior work (Park et al., 2025; Chen et al., 2024). Ground-truth correctness labels are provided by GPT-4o-mini (Hurst et al., 2024), which is in-

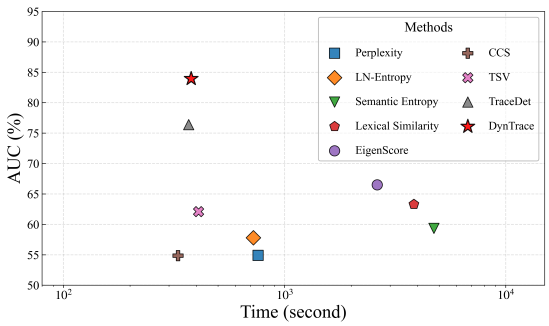


Figure 4: Pareto-frontier of performance vs. efficiency.

Table 2: Zero-shot cross-task generalization matrix.

Method	TriviaQA		HotpotQA		CSQA		Avg
	H-QA	CSQA	T-QA	CSQA	T-QA	H-QA	
Latent-based							
CCS	50.1	54.5	51.8	54.0	54.2	56.6	58.2
TSV	58.5	65.3	65.5	59.1	56.2	63.2	61.3
Trajectory-based							
TraceDet	73.1	61.5	57.4	65.0	**74.8**	66.2	66.3
DynHD	**85.5**	**68.7**	**73.3**	**64.9**	73.6	**71.4**	**72.9**

Table 3: Ablation study in AUROC(%).

Ablation Variant	TriviaQA		HotpotQA		CSQA	
	128	64	128	64	128	64
LLaDA-8B-Instruct						
DynHD (Full Framework)	**86.7**	**86.1**	**84.2**	**85.3**	81.3	**84.3**
Semantic-aware Evidence Construction Components						
w/o Token Filtering	74.8	73.6	75.5	78.1	74.1	76.6
w/o Mean Uncertainty (μ_{avg})	78.6	81.1	79.0	80.2	78.5	77.1
w/o Max Entropy (s_{max})	82.5	83.1	81.8	82.3	79.1	82.1
w/o Top-k Entropy ($\mu_{\text{top-}k}$)	82.7	82.6	81.7	83.0	78.9	82.5
Dynamical Deviation Learning Components						
Naive Average	57.3	65.5	51.7	48.3	63.2	61.8
Flattened Temporal Aggregation	62.3	70.5	66.8	57.5	67.1	66.2
Uniform Temporal Pooling (w/o Attn)	84.5	84.8	82.4	83.1	79.5	80.5
w/o $\mathcal{L}_{path}$ (Stagnation)	84.1	85.3	84.1	83.8	**81.6**	83.2
w/o $\mathcal{L}_{reb}$ (Rebound)	85.2	85.6	**84.2**	83.5	80.6	83.3
Dream-7B-Instruct						
DynHD (Full Framework)	**87.3**	**84.4**	**80.1**	**85.6**	**83.5**	84.6
Semantic-aware Evidence Construction Components						
w/o Token Filtering	75.4	72.1	69.8	74.2	73.6	73.9
w/o Mean Uncertainty (μ_{avg})	79.2	78.5	74.5	79.1	76.2	77.8
w/o Max Entropy (s_{max})	83.1	81.2	77.5	82.3	80.9	81.5
w/o Top-k Entropy ($\mu_{\text{top-}k}$)	83.5	80.9	77.8	82.6	80.5	81.9
Dynamical Deviation Learning Components						
Naive Average	58.2	59.1	48.5	52.3	60.1	57.8
Flattened Temporal Aggregation	65.1	66.3	59.4	62.7	68.2	65.6
Uniform Temporal Pooling (w/o Attn)	84.5	82.1	78.0	83.1	81.2	81.8
w/o $\mathcal{L}_{path}$ (Stagnation)	85.2	83.5	79.2	84.5	82.8	83.8
w/o $\mathcal{L}_{reb}$ (Rebound)	86.1	83.1	79.5	84.2	82.7	**84.8**

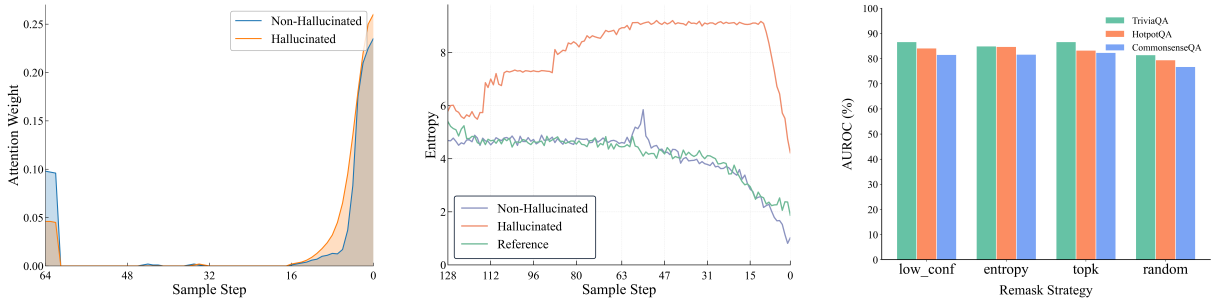
structed to judge whether a generated answer is factually correct based on its consistency with the reference answer and its own internal knowledge (Zheng et al., 2023; Liu et al., 2023). We further assess the reliability of the automatic annotation by measuring agreement with human judgments, achieving an agreement rate of 94%. We partition the dataset into 1,400 samples for training, 300 for validation, and 300 for testing. Model selection is based on validation AUROC, and all reported results are obtained on the held-out test set. We provide detailed descriptions of the baselines, benchmarks, and experimental settings in Appendix C, with additional experiments in Appendix D.

4.2 Experimental Results

Main Results. Table 1 presents a comprehensive comparison of DynHD against baseline methods across two D-LLMs and three factuality QA datasets with varying generation lengths. From the results, we have several observations. ❶ DynHD achieves the highest performance in almost all settings. On the LLaDA-8B-Instruct model, it outperforms the strongest baseline by an average of 12.2% absolute AUROC, while on Dream-7B-Instruct, it reaches an average of 84.3%. These consistent gains exhibit the value of exploiting fine-grained deviation dynamics and convergence morphology rather than relying solely on output uncertainty or static hidden-state representations. ❷ Output-based methods lack consistency across datasets, showing poor performance on several cases. This occurs because multiple-choice answers are often confident but wrong, rendering uncertainty-based metrics ineffective. ❸ Latent-based approaches like EigenScore and TSV show more competitive results on Dream-7B-Instruct by capturing hidden-state geometry, but remain sensitive to dataset shifts.

Inference Efficiency. Figure 4 shows the balance between performance and efficiency. DynHD occupies the upper-left quadrant, representing the best results in our evaluation. While its speed is comparable to fast baselines like CCS, DynHD achieves significantly higher accuracy. Notably, our approach is much more efficient than multi-sample methods like Semantic Entropy, which often have high computational costs caused by multi-round sampling. Overall, DynHD provides an ideal balance of accuracy and speed, allowing for real-time monitoring without causing extra delays.

Cross-dataset Generalization Ability. To evaluate robustness, we conduct zero-shot cross-dataset assessments in table 2, where DynHD yields the highest average AUROC (72.9%). It notably achieves a 85.5% AUROC in TriviaQA to H-QA transfer, significantly outperforming TraceDet (73.1%) and suggesting that our dynamical encoding captures universal denoising laws across domains. On the challenging CSQA source, DynHD



(a) (b) (c)

Figure 5: (a) Visualization of attention weights ω_t on different steps. (b) Top-k entropy trajectories of representative samples on HotpotQA ($T = 128$). (c) Performance comparison of different D-LLMS remasking strategies.

remains robust (e.g., 73.6% on T-QA), despite it slightly under-performs TraceDet. This is because the fixed options lead to a faster decision process, which reduces the observable uncertainty signals.

Ablation Study. As summarized in Table 3, we evaluate the contributions of key designs in DynHD, coming with the following findings. In terms of the evidence construction module, ❶ without token filtering, the performance drops significantly, which proves that filtering out noise is essential to get clean signals. ❷ Removing individual statistical evidences consistently degrades results, proving that robust detection requires simultaneous monitoring of semantic stability, localized anomaly, and regional instability. For the deviation learning module, ❸ *Naive Average* variant yields poor performance because simple arithmetic averaging cannot effectively distinguish hallucinations. ❹ *Flattened Temporal Aggregation* underperforms due to its lack of explicit temporal dynamic analysis. ❺ Removing attention-based weighting degrades performance, as the model fails to capture critical dynamic changes, especially at the late stages of the denoising process. ❻ Our regularization terms further improve the model's ability to differentiate hallucinations from factual generations by capturing and penalizing specific non-convergent trajectories, such as *Stagnation* and *Rebound* patterns.

Visualization of Temporal Weighting. Figure 5a shows the attention weights for LLaDA-8B-Instruct on the TriviaQA dataset ($t = 64$). We can see the learned weights are mostly concentrated in the final steps of the denoising process. This matches our observation that stagnation and rebound patterns usually happen at the end of the trajectory. By focusing on these late steps, our model can find clear signs of hallucinations and ignore early random noise.

Visualization of Reference Trajectories. Figure 5b shows entropy trajectories on the HotpotQA dataset. We introduce the learned reference path to show how a correct denoising process should look. The factual sample stays close to this path and drops smoothly toward zero. In contrast, a hallucinated sample is unstable and shows a clear rebound at the end. This shows that the shape of the curve is more important than its absolute value. The introduction of reference generator makes it easy to find denoising processes with hallucination.

Robustness to Masking Strategies. Figure 5c examine the performance of DynHD on LLaDA-8B-Instruct under various internal remasking strategies employed by D-LLMs during generation. The results across multiple datasets consistently demonstrate competitive detection performance, regardless of the specific spatial filtering logic used (such as entropy-based or confidence-based remasking) (Ghazvininejad et al., 2019; Savinov et al., 2022). This stability suggests that the proposed framework is adaptable and maintains its effectiveness across diverse D-LLM configurations.

5 Conclusion

We propose DynHD to detect D-LLM hallucinations using trajectory-based deviation modeling. DynHD identifies factual errors by comparing actual evidence against a learned reference. This approach effectively captures key signals such as late-stage stagnation and uncertainty rebound. Experiments on mainstream models show that DynHD achieves superior performance with high computational efficiency. Overall, this work demonstrates that denoising dynamics can serve as reliable signals for hallucination detection of D-LLMs.

Limitations

Although DynHD demonstrates strong performance on hallucination detection across standard QA benchmarks, the evaluation scope remains limited. To better assess the generality of our approach, future work should explore more diverse task domains, such as long-form summarization and open-ended generation, where factual boundaries and structural noise are more complex. Additionally, the current design relies on training an evidence dynamics generator using in-domain factual samples to establish reference trajectories. This dependency may limit its zero-shot adaptability in entirely unfamiliar or out-of-distribution scenarios. Furthermore, since our framework is intrinsically tailored to the continuous denoising process of Diffusion LLMs, adapting these trajectory-based insights to standard auto-regressive models remains an open challenge.

Ethics Statement

Our research focuses exclusively on scientific questions, with no involvement of human subjects, animals, or environmentally sensitive materials. Therefore, we foresee no ethical risks or conflicts of interest. We are committed to maintaining the highest standards of scientific integrity and ethics to ensure the validity and reliability of our findings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*.

Shenxu Chang, Junchi Yu, Weixing Wang, Yongqiang Chen, Jialin Yu, Philip Torr, and Jindong Gu. 2026. Tracedet: Hallucination detection from the decoding trace of diffusion large language models. In *International Conference on Learning Representations*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Xuefeng Du, Chaowei Xiao, and Sharon Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6112–6121.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2023. Diffuseq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, and 1 others. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yatai Ji, Teng Wang, Yuying Ge, Zhiheng Liu, Sidi Yang, Ying Shan, and Ping Luo. 2025. From denoising to refining: A corrective framework for vision-language diffusion model. *arXiv preprint arXiv:2510.19871*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024. On large language models' hallucination with regard to known facts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Gen Li and Changxiao Cai. 2025. Breaking ar's sampling bottleneck: Provable acceleration via diffusion language models. In *Advances in neural information processing systems*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. 2025. A survey on diffusion language models. *arXiv preprint arXiv:2508.10875*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models, 2024. *URL https://arxiv. org/abs/2305*, 19187.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *Advances in Neural Information Processing Systems*.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 1938–1952.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. 2025. Steer llm latents for hallucination detection. In *International Conference on Machine Learning*.

Elizabeth Pavlova and Xue-Xin Wei. 2025. Diffusion models under low-noise regime. *arXiv preprint arXiv:2506.07841*.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. 2022. Step-unrolled denoising autoencoders for text generation. In *International Conference on Learning Representations*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wen Wang, Bozhen Fang, Chenchen Jing, Yongliang Shen, Yangyi Shen, Qiuyu Wang, Hao Ouyang, Hao Chen, and Chunhua Shen. 2026. Time is a feature: Exploiting temporal dynamics in diffusion language models. In *International Conference on Learning Representations*.

Xu Wang, Chenkai Xu, Yijie Jin, Jiachun Jin, Hao Zhang, and Zhijie Deng. 2025. Diffusion llms can do faster-than-ar inference via discrete diffusion forcing. *arXiv preprint arXiv:2508.09192*.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. Interrogatellm: Zero-resource hallucination detection in llm-generated answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. 2023. Do large language models know what they don't know? In *Findings of the association for Computational Linguistics: ACL 2023*, pages 8653–8665.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Related Work

Hallucination Detection Most existing hallucination detection studies focus on Autoregressive Language Models (AR-LLMs) (Zhang et al., 2025; Ji et al., 2023). These methods typically extract signals from the final prediction, such as token-level uncertainty measures (e.g., entropy or perplexity (Malinin and Gales, 2020; Ren et al., 2023)), or assess semantic consistency across multiple samples (e.g., semantic entropy (Kuhn et al., 2023; Manakul et al., 2023)). Some recent work also probes internal representations during the forward pass (e.g., hidden states (Azaria and Mitchell, 2023; Li et al., 2023; Marks and Tegmark, 2024)) to infer factuality. However, diffusion large language models (D-LLMs) differ fundamentally from AR-LLMs in their generation mechanism: hallucination cues are not confined to the final step but are distributed across a denoising trajectory spanning the entire iterative process. This process-centric nature makes static, outcome-based detectors prone to missing early discriminative evidence exposed along the trajectory (Chuang et al., 2024).

Diffusion Large Language Models Diffusion Large Language Models (D-LLMs) extend the remarkable success of diffusion paradigms (Yang et al., 2023; Wang et al., 2025) to the discrete text domain (Li et al., 2022). By unifying the discrete remasking process, LLaDA has successfully scaled D-LLMs to the 8B parameter level, achieving performance parity with leading AR-LLMs models such as LLaMA-3 (Grattafiori et al., 2024). Furthermore, Dream-7B demonstrates the architectural versatility of this paradigm by adopting the identical configurations of Qwen2.5-7B (Hui et al., 2024) and training it under a diffusion objective. On the reliability side, Wang et al. (2026) first identified the phenomenon of temporal oscillation, where correct answers may emerge at intermediate denoising steps but can be overridden by erroneous late-stage refinements. Building on the trajectory view, TraceDet (Chang et al., 2026) proposes a trajectory-based detection framework that models generation as an action trace and leverages the Information Bottleneck principle to automatically select the most discriminative sub-trace steps (Paranjape et al., 2020; Tishby and Zaslavsky, 2015). **DynHD** builds upon these process-based insights but moves beyond simplistic sub-trace selection by explicitly modeling the fine-grained denoising dynamics and mitigating spatial padding noise to achieve superior hallucination detection.

Denoising Dynamics Denoising dynamics characterizes how a model progressively removes randomness and approaches the target distribution through multi-step refinement (Song et al., 2021). Li and Cai (2025) provide theoretical analysis suggesting that, under ideal conditions, the sampling error of D-LLMs should decrease as the number of denoising steps increases and eventually converge to the data distribution. This implies that reliable generation trajectories should be accompanied by smooth uncertainty dissipation. When hallucinations arise due to insufficient knowledge (Yin et al., 2023; Jiang et al., 2024), this dynamical balance can break: uncertainty may remain persistently high because of unresolved inconsistencies, leading to stagnation, or exhibit abnormal rebound in mid-to-late stages when refinement fails (Su et al., 2024; Huang et al., 2025). Such morphological deviations from a normal convergence path directly reflect instability in the generation process (Pavlova and Wei, 2025). This work is the first to utilize such continuous temporal evolution patterns for D-LLM hallucination detection, advancing the field from static uncertainty summaries toward a comprehensive process-level characterization of generative reliability.

B Algorithmic Details of DynHD

B.1 Pseudocode

Algorithm 1 details the complete process of our algorithm.

B.2 Complexity Analysis

The computational overhead of DynHD is negligible compared to the D-LLM inference cost, as it operates on pre-computed uncertainty fields without re-invoking the backbone.

- **Spatial Filtering:** For each step t, position-wise masking is $\mathcal{O}(l)$, and extracting Top-k statistics requires $\mathcal{O}(l \log k)$ using a min-heap. The total complexity for a full trajectory is $\mathcal{O}(T \cdot l \log k)$. Since the sequence length l is typically much smaller than the vocabulary size $|\mathcal{V}|$, this process is significantly faster than the standard logit-to-probability conversion in Transformer heads.

- **Temporal Modeling:** Our temporal encoder and attention mechanism operate on the condensed 3D evidence vectors. With a hidden dimension h,

Algorithm 1 The two-stage training algorithm of DynHD

Require: Training pool $\mathcal{X}$; ignore set $\mathcal{I}_{\text{ignore}}$; hyperparameter k; epochs E_1, E_2; batch size B.
Ensure: Optimized reference parameters θ and detector parameters Φ.

// Stage 1: Evidence Dynamics Generator (Sec. 3.2)
Initialize generator parameters θ;
for $epoch = 1$ **to** E_1 **do**
  Sample mini-batch $\{(\mathcal{T}, \mathbf{q}, y)_i\}_{i=1}^{B} \subset \mathcal{X}$ where $y = 0$ {Factual samples only}
  for all $(\mathcal{T}, \mathbf{q})$ **in** mini-batch **do**
    Evidence Construction: Extract valid semantic positions $\mathcal{K}_t$ via $\mathcal{I}_{\text{ignore}}$;
    Condense step-wise uncertainty into $\mathbf{a}_t = [\mu_{\text{keep}}^{(t)}, s_{\max}^{(t)}, \mu_{\text{top-}k}^{(t)}]$ (Eq. 4);
    Predict reference trajectory $\hat{\mathbf{a}}_t = g_\theta(\mathbf{q}, \mathbf{e}_t)$ using query and timestep embeddings;
  end for
  Update θ by minimizing $\mathcal{L}_{\text{ref}}$ via SmoothL1 loss (Eq. 5);
end for
// Stage 2: Deviation-based Hallucination Detector (Sec. 3.2)
Initialize detector parameters Φ and adaptive factual margins $\hat{m}^{(\tau)}$;
for $epoch = 1$ **to** E_2 **do**
  Sample mini-batch $\{(\mathcal{T}, \mathbf{q}, y)_i\}_{i=1}^{B} \subset \mathcal{X}$;
  for all $(\mathcal{T}, \mathbf{q}, y)$ **in** mini-batch **do**
    Evidence Construction: Obtain actual evidence trajectory $\mathcal{E} = (\mathbf{a}_T, \ldots, \mathbf{a}_0)$ from $\mathcal{T}$ via Eq. 1 & 2;
    Generate reference trajectory $\hat{\mathcal{E}} = (\hat{\mathbf{a}}_T, \ldots, \hat{\mathbf{a}}_0)$ via frozen g_θ;
    Compute velocity $\Delta\mathbf{a}_t = \mathbf{a}_{t-1} - \mathbf{a}_t$ to capture transition dynamics;
    Construct composite feature $\mathbf{x}_t = [\mathbf{a}_t; \hat{\mathbf{a}}_t; \Delta\mathbf{a}_t]$ for each step t;
    Project $\mathbf{x}_t$ to $\mathbf{z}_t$, compute weight ω_t (Eq. 6), and aggregate to $\mathbf{h}$;
    Compute path-deviation score s_{path} (Eq. 7) and rebound score s_{reb} (Eq. 8);
  end for
  Update adaptive margins $\hat{m}^{(\tau)}$ via EMA using scores of factual samples ($y = 0$) in the batch (Eq. 9);
  Update Φ by minimizing total loss $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_{\text{path}} + \lambda_2\mathcal{L}_{\text{reb}}$ (Eq. 10);
end for

the complexity is $\mathcal{O}(T \cdot (d \cdot h + h^2))$, where $d = 3$. Since d and h are small constants independent of the LLM's scale, the inference latency of the detection head is essentially constant, adding no perceptible delay to the generation pipeline.

As a result, DynHD maintains an efficient $\mathcal{O}(T \cdot l)$ footprint, which is dwarfed by the quadratic self-attention complexity $\mathcal{O}(T \cdot l^2)$ of the underlying D-LLM.

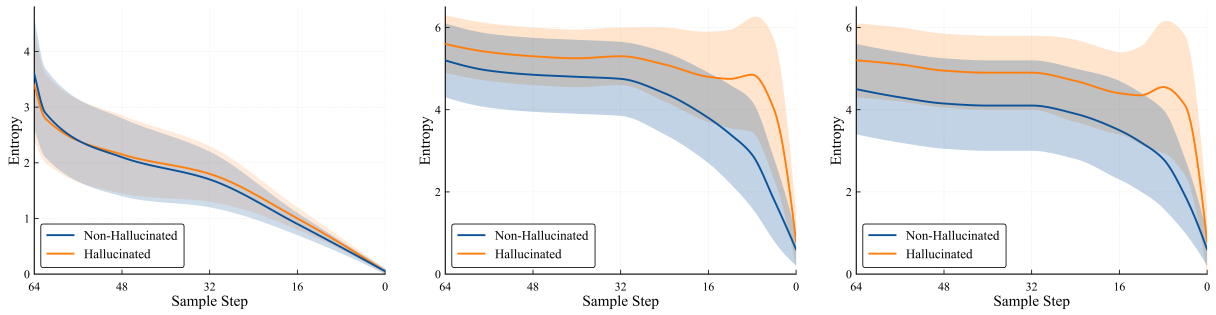
C Detailed Experimental Setup

C.1 Datasets and Benchmarks

To ensure a rigorous and multi-faceted evaluation, we conduct experiments across three representative factuality-oriented QA benchmarks, following the protocol established in recent trajectory-based studies (Chang et al., 2026):

- **TriviaQA** (Joshi et al., 2017): A large-scale open-domain dataset focusing on precise entity retrieval and factual knowledge.
- **HotpotQA** (Yang et al., 2018): A multi-hop reasoning benchmark that requires synthesizing information across multiple context fragments to maintain logical consistency.
- **CommonsenseQA** (Talmor et al., 2019): A task evaluating general reasoning capabilities, where the model must navigate commonsense logic beyond surface-level pattern matching.

All dataset preprocessing, including prompt templates and data splits, strictly adheres to the configurations used in the comparative studies to ensure a level playing field.



(a) Global Distinguishability (b) HotpotQA Results (c) TriviaQA Results

Figure 6: Empirical analysis of the distinguishability of statistical evidence trajectories. (a) shows the enhanced separation after structure-aware filtering, while (b) and (c) demonstrate the consistent morphological divergence between factual and hallucinatory samples across datasets.

C.2 Baseline Configurations

We compare DynHD against seven established baselines, categorized into two primary paradigms based on their information source:

(1) Output-Centric Approaches: These methods detect hallucinations by analyzing the generated text or its output probability distribution:

- *Perplexity* (Ren et al., 2023): Identifies potential hallucinations based on the negative log-likelihood assigned by the model to its own sequence.
- *LN-Entropy* (Malinin and Gales, 2020): Employs length-normalized predictive entropy to flag generations with unusually high average uncertainty.
- *Semantic Entropy* (Kuhn et al., 2023): Quantifies consistency by partitioning multiple stochastic outputs into semantic equivalence classes.
- *Lexical Similarity* (Lin et al., 2024): Assesses hallucinations via surface-level lexical overlap across independent generation paths.

(2) Latent-Centric Approaches: These methods probe the internal representations and hidden states of the transformer architecture:

- *EigenScore* (Chen et al., 2024): Evaluates response consistency through the covariance structure of internal latent embeddings.
- *CCS* (Burns et al., 2023): Searches for a contrastive-consistent direction in the latent space to distinguish truthful from untruthful activations.
- *TSV* (Park et al., 2025): Utilizes a separator vector to classify truthfulness based on learned centroids within the latent space.

Finally, we include **TraceDet** (Chang et al., 2026) as the primary trajectory-based baseline. TraceDet characterizes the denoising path using a uniform temporal aggregation strategy, providing a representative point of comparison for analyzing the evolution of uncertainty during the denoising process.

C.3 Implementation Details

Our experiments, including model inference and detector training, are conducted on **NVIDIA RTX 3090 GPUs (24GB VRAM)**.

C.4 Hyperparameter Search Space

We conduct a grid search to optimize the performance of DynHD. The variables and their corresponding search ranges are summarized in Table 4.

Table 4: Search space of key hyperparameters in DynHD.

Variable	Search Range
$\mathrm{LR}_{\mathrm{stage1}}$	$\{3 \times 10^{-4}, 1 \times 10^{-3}\}$
$\mathrm{LR}_{\mathrm{stage2}}$	$\{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}\}$
$\mathrm{WD}_{\mathrm{stage1}}$	$\{0, 0.01, 0.05\}$
$\mathrm{WD}_{\mathrm{stage2}}$	$\{0, 1 \times 10^{-4}, 1 \times 10^{-3}\}$
λ_1 (Path)	$[0.0, 0.4]$ (step 0.05)
λ_2 (Rebound)	$[0.0, 0.4]$ (step 0.05)
$\lambda_{\mathrm{warmup}}$	$\{0, 0.3\}$

D Additional Experimental Results

D.1 Case of Hallucination Dynamics

Table 5 details the four samples discussed in our Introduction. For factual cases like the New York projects and the Shining Path, the denoising entropy steadily decreases as the model converges on the correct facts in the late-stage. In contrast, hallucinations show clear instability. In the Disney case,

Dataset	Type	Question & Response
HotpotQA	Factual	**Q:** Are Manhattan West and Singer Building both projects in New York? **R:** Yes.
	Hallucination	**Q:** What State has a Disney Resort & Spa that is a beachside hotel, resort and vacation destination offering complimentary children's activities and programs and that Djuan Rivers was a General Manager at? **R:** The Disney Resort and Spa where Djuan Rivers served as General Manager is located in Florida.
TriviaQA	Factual	**Q:** Where did the Shining Path terrorists operate? **R:** The Shining Path terrorists operated in Peru.
	Hallucination	**Q:** Who had a 70s No 1 hit with Kiss You All Over? **R:** The Village People had a 70s No 1 hit with "Kiss You All Over."

Table 5: Examples of factual and hallucinated samples. Each dataset displays a Factual sample followed by a Hallucination. Factual samples exhibit consistent convergence, whereas hallucinations trigger detectable instability in denoising dynamics.

the model is biased toward "Florida" maybe due to the keyword "Disney," but the specific constraint "Djuan Rivers" causes an entropy rebound when the model fails to resolve the conflict. Similarly, the hallucination for the 70s hit shows a late-stage bounce, indicating a breakdown in factual retrieval. These examples confirm that the trajectory shape is a more reliable signal for detection than absolute entropy values.

D.2 Distinguishability of Statistical Evidence Trajectories

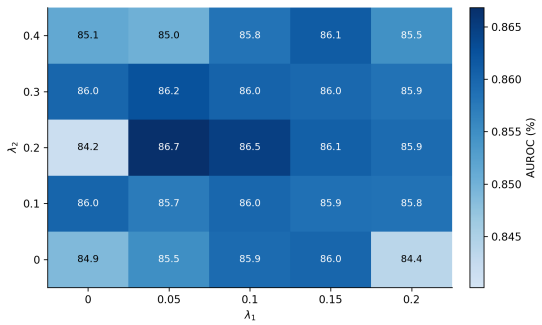


Figure 7: AUROC heatmap for varying λ_1 (Path) and λ_2 (Rebound) on TriviaQA. The peak region validates the synergistic effect of both dynamical regularizers.

We test how our metrics separate factual answers from hallucinations in Figures 6a–6c. As shown in Figure 6a, our filtering method clarifies the uncertainty data. This creates a clear difference between facts and hallucinations, whereas raw data often overlaps. Specific tests in Figures 6b and 6c show a common pattern: factual paths smoothly go down, while hallucinations bounce at the end.

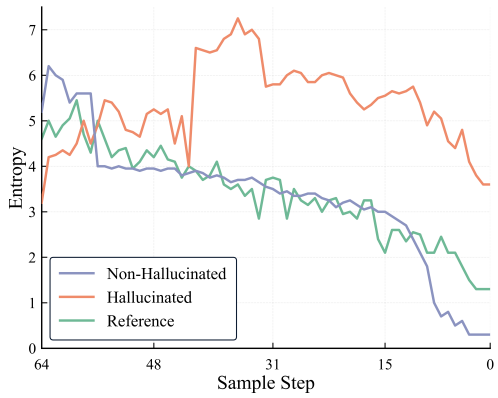


Figure 8: Empirical denoising dynamics of a_t (observed) and $\hat{a}_t$ (reference) at 64 steps on HotpotQA.

D.3 Sensitivity of λ_1 and λ_2

Figure 7 shows how λ_1 (Path) and λ_2 (Rebound) work together on TriviaQA. The heatmap indicates that the model works best when both λ_1 and λ_2 are used (around 0.1 to 0.3). This confirms that tracking the whole path and checking the late-stage bounce are both important. The model performs well across a wide range of values, showing it is not overly sensitive to specific parameter settings.

D.4 Denoising Dynamics at 64 Steps

Figure 8 compares the real paths a_t and the reference curves $\hat{a}_t$ at 64 steps. Factual answers follow the reference curve closely, but hallucinations show a clear bounce. These results prove that the shape of the curve is the most important indicator of an error, regardless of the exact values.