Toward Consistent World Models with Multi-Token Prediction and Latent Semantic Enhancement

Qimin Zhong[1], **Hao Liao**[1,2], **Haiming Qin**[1], **Mingyang Zhou**[1],
Rui Mao[1], **Wei Chen**[3], **Naipeng Chao**[2,4]
[1]College of Computer Science and Software Engineering, Shenzhen Univerisity
[2]Provincial Key Laboratory of Intelligent Communication and Digital Society Governance, Shenzhen University
[3]Microsoft Research Asia
[4]School of Media and Communication, Shenzhen Univerisity
{2023044007@email, haoliao, 2453103002@mails, zmy, mao, npchao}@szu.edu.cn
weic@microsoft.com

Abstract

Whether Large Language Models (LLMs) develop coherent internal world models remains a core debate. While conventional Next-Token Prediction (NTP) focuses on one-step-ahead supervision, Multi-Token Prediction (MTP) has shown promise in learning more structured representations. In this work, we provide a theoretical perspective analyzing the gradient inductive bias of MTP, supported by empirical evidence, showing that MTP promotes the convergence toward internal belief states by inducing representational contractivity via gradient coupling. However, we reveal that standard MTP often suffers from structural hallucinations, where discrete token supervision encourages illegal shortcuts in latent space that violate environmental constraints. To address this, we propose a novel method **Latent Semantic Enhancement MTP (LSE-MTP)**, which anchors predictions to ground-truth hidden state trajectories. Experiments on synthetic graphs and real-world Manhattan Taxi Ride show that LSE-MTP effectively bridges the gap between discrete tokens and continuous state representations, enhancing representation alignment, reducing structural hallucinations, and improving robustness to perturbations.

1 Introduction

Internalizing the dynamics of an environment is a hallmark of intelligent behavior. This capability, often formalized as a world model (Ha and Schmidhuber, 2018; Schmidhuber, 1990), allows an agent to reason beyond immediate observations by simulating how states evolve over time (Silver et al., 2017; Schrittwieser et al., 2019). Rather than reacting myopically to inputs, systems equipped with world models can anticipate future outcomes, evaluate alternative trajectories, and plan accordingly. The success of DreamerV3 (Hafner et al., 2025) vividly illustrates how learning internal dynamics can yield strong generalization across diverse tasks, even under limited supervision. Recent evidence suggests that the fidelity of internal world models is a key driver of post-training potential and correlates with improved reasoning and downstream performance (Gupta et al., 2025).

In the context of Natural Language Processing, this perspective raises a fundamental and intriguing question: do Large Language Models (LLMs) trained purely through Next-Token Prediction (NTP) develop meaningful internal world models (Brown et al., 2020; Rae et al., 2021)? While NTP has proven remarkably effective at scaling language understanding and generation, its optimization objective is inherently local, as it primarily focuses on predicting the likelihood of the next symbol given a context. As a result, such models often excel at capturing surface-level regularities but struggle to consistently internalize deeper global structure or long-range dynamics, especially when complex reasoning requires maintaining coherent latent states over extended horizons (Bachmann and Nagarajan, 2024; Wyatt et al., 2025).

This concern has been substantiated by recent real-world evaluations. Vafa et al. (2024) introduce a world-model benchmark based on Manhattan taxi trajectories, where city streets are abstracted as a graph with explicit topological constraints. Despite achieving near-perfect next-step prediction accuracy, NTP-trained models frequently fail to encode the global structure of the street network in their latent states, leading to invalid routes and severe fragility under minor perturbations. These findings demonstrate that strong token-level performance alone does not guarantee a coherent internal world model.

Multi-Token Prediction (MTP) has recently emerged as a promising alternative (Gloeckle et al., 2024). By supervising multiple future tokens simultaneously, MTP encourages models to look beyond immediate continuations and consider longer-term evolution. This shift in supervision fundamentally

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13582–13602

alters the training signal: instead of fitting isolated conditional distributions, the model is pressured to represent how sequences unfold over time. From a representation-learning standpoint, such foresight can induce representational contractivity, encouraging diverse historical contexts to converge toward shared internal belief states that summarize the underlying environment. This phenomenon suggests a potential pathway for LLMs to move from shallow sequence modeling toward more structured internal representations resembling world models.

Yet, the presence of foresight alone does not guarantee coherent internal reasoning. In practice, we observe that MTP-trained models can develop a subtle but systematic failure mode, which we refer to as structural hallucination. Even when long-term predictions are accurate at the token level, the latent evolution that supports them may violate essential constraints of the environment. Intermediate steps can be implicitly skipped, transitions may become implausible, and internal trajectories can exploit shortcuts that would be invalid under the true dynamics. This reveals a key tension: optimizing distant predictions without explicit trajectory-level grounding can incentivize models to prioritize outcomes over the integrity of the underlying process.

These observations point to a broader gap between discrete supervision and continuous internal dynamics. Token-level objectives, even when extended to multiple future steps, offer limited control over how representations evolve over time. In the absence of mechanisms that explicitly align latent transitions with valid state progressions, models may develop internally inconsistent simulations that appear accurate only at their final predictions. Bridging this gap is crucial for elevating multi-token prediction from a stronger forecasting objective to a dependable foundation for world modeling and long-horizon planning.

This work relates to several active research threads, including world models in language modeling, multi-token prediction, latent state consistency, and graph-based planning. Detailed discussion is deferred to Appendix A.

To summarize, our main contributions are highlighted by the following three perspectives:

- We provide a theoretical analysis of the gradient coupling mechanism in Multi-Token Prediction (MTP), showing how it induces contractivity that facilitates the emergence of belief states, while exposing a structural hallucination risk arising from overemphasis on distant targets over local connectivity.
- We propose LSE-MTP, a framework that enforces latent consistency by aligning multi-token predictions with ground-truth hidden state trajectories and semantic anchors, thereby enforcing valid stepwise transitions and discouraging illegal shortcuts.
- Through extensive experiments on synthetic graphs and real-world Manhattan taxi navigation, we show that LSE-MTP improves path legality, belief compression, and robustness to perturbations in multi-step planning.

2 Preliminaries

2.1 Next-Token Prediction

The standard paradigm for autoregressive sequence modeling is Next-Token Prediction (NTP). Given a history $H_n = (u_1, \ldots, u_n)$, the objective minimizes the negative log-likelihood of the next token:

$$\mathcal{L}_{\text{NTP}}(\theta) = \mathbb{E}_{S\sim\mathcal{D},n}\big[-\log P_\theta(u_{n+1} \mid H_n)\big]. \quad (1)$$

Despite its empirical success, NTP exhibits limitations in structured reasoning tasks: (i) it primarily fits local co-occurrence statistics rather than invariant transition rules (Wu et al., 2024), and (ii) under teacher forcing, models can exploit local token correlations to bypass global reasoning, leading to the acquisition of shortcuts during training that fail to generalize to the underlying task logic (Bachmann and Nagarajan, 2024; Frydenlund, 2025).

2.2 Multi-Token Prediction

Multi-Token Prediction (MTP) extends NTP by jointly predicting the next K future tokens during training, while retaining standard autoregressive decoding at inference (Gloeckle et al., 2024).

We consider an MTP architecture with a shared output head and horizon-specific transition layers. Given the backbone hidden state $\mathbf{h}_n = f_\theta(H_n)$, the next token u_{n+1} is predicted directly, while the k-step future token ($k \geq 2$) is predicted from a transformed representation $\mathcal{T}_\phi^{(k-1)}(\mathbf{h}_n)$. All predictions for different horizons are decoded by the same shared output head. The training objective is:

$$\mathcal{L}_{\text{MTP}} = \mathbb{E}_{S,n}\Big[\mathcal{L}^{(1)}(\mathbf{h}_n, u_{n+1}) + \sum_{k=2}^{K} \mathcal{L}^{(k)}(\mathbf{h}_n, u_{n+k})\Big]. \quad (2)$$

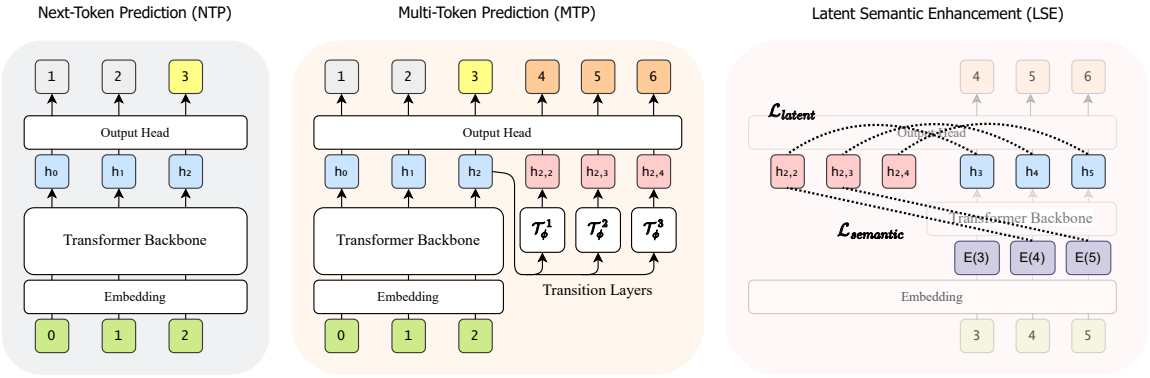

Figure 1: **Overview of LSE-MTP.** Given a backbone hidden state $\mathbf{h}_n$, horizon-specific transition layers produce multi-step predictive representations. Training combines multi-step token prediction with latent consistency and semantic anchoring losses. All transition layers are discarded at inference time.

2.3 Representation Space and Belief States

The hidden state $\mathbf{h}_n$ serves as a compact summary of the history H_n and implicitly encodes information about future trajectories.

Definition 1 *The set of hidden states $\mathcal{H} = \{\mathbf{h}_n\}$ forms a **representation space**, where histories with similar future continuations are embedded nearby (Littman and Sutton, 2001).*

Definition 2 *The idealized representation associated with $\mathbf{h}_n$ is a **belief state** $\mathbf{b}_n$ (Kaelbling et al., 1998), satisfying*

$$P(u_{n+1:\infty} \mid H_n) \approx P(u_{n+1:\infty} \mid \mathbf{b}_n). \quad (3)$$

Belief states provide a compact internal model of future dynamics.

3 A Theoretical Perspective on Multi-Token Prediction

Before diving into mathematical analysis, we provide the intuition behind MTP's impact. By predicting multiple tokens simultaneously, MTP encourages histories leading to the same future to "merge" within the representation space. This merging is inherently blind: it constrains only future outcomes while ignoring intermediate states, which can produce illegal shortcuts in latent space. In this section, we formally characterize this behavior using gradient flow dynamics.

To obtain a tractable analytic framework, we focus on the **linearized regime** (lazy training), approximating the optimization trajectory via the local Neural Tangent Kernel (NTK) (Chizat et al., 2019). This local linearization captures the instantaneous directional pressure exerted by the loss on the representation space.

Let $\mathbf{h} = f_\theta(H)$ denote the hidden state of a backbone parameterized by θ, evolving under gradient flow $\dot{\theta} = -\eta \nabla_\theta \mathcal{L}$. We define the representation-level NTK as:

$$\mathbf{K}(\mathbf{h}_i, \mathbf{h}_j) = \nabla_\theta f_\theta(H_i) \nabla_\theta f_\theta(H_j)^\top \in \mathbb{R}^{d \times d}.$$

Definition 3 *Two hidden states $\mathbf{h}_1$ and $\mathbf{h}_2$ are **k-step future equivalent** ($\mathbf{h}_1 \sim_k \mathbf{h}_2$) if they are supervised by the same k-step-ahead target token $y^* = u_{n+k} = u_{m+k}$ under the $(k-1)$-th transition layer $\mathcal{T}^{(k-1)}$ and the shared prediction head.*

Definition 4 *The representation space exhibits **contractivity** for a pair of histories if the time derivative of the squared distance $\mathcal{D}(\mathbf{h}_1, \mathbf{h}_2) = \|\mathbf{h}_1 - \mathbf{h}_2\|^2$ satisfies $\dot{\mathcal{D}} \leq 0$ under gradient flow, indicating convergence toward a unified belief state.*

Based on these definitions, we compare the geometric effects of NTP and MTP. Formal derivations are deferred to Appendix B.

Theorem 1 *Under the NTP loss $\mathcal{L}_{NTP}$, the contractive condition $\dot{\mathcal{D}} \leq 0$ holds primarily for 1-step equivalent states ($\mathbf{h}_1 \sim_1 \mathbf{h}_2$). For states with different next-step targets, the gradients $\nabla_{\mathbf{h}} \mathcal{L}$ tend to point in opposite directions, preserving representational separation.*

Theorem 2 *Under the MTP loss $\mathcal{L}_{MTP}$, consider k-step future-equivalent states $\mathbf{h}_1 \sim_k \mathbf{h}_2$ with different immediate targets $u_{n+1} \neq u_{m+1}$. A k-step update on $\mathbf{h}_1$ induces a positive cross-update on the corresponding logit of $\mathbf{h}_2$, $\dot{z}_{y_1}(\mathbf{h}_2) > 0$, where the gradients $\nabla_{\mathbf{h}_1} \mathcal{L}_1^{(k)}$ and $\nabla_{\mathbf{h}_2} \mathcal{L}_1^{(k)}$ align through the cross-history NTK $\mathbf{K}(\mathbf{h}_1, \mathbf{h}_2)$, facilitating a predictive coupling that can partially blur the representational separation between distinct trajectories.*

Intuition: If two histories share an identical future, training on one trajectory inadvertently increases the prediction confidence of the other's next token, even if their immediate targets differ.

Lemma 1 *For a pair of k-step future-equivalent states ($\mathbf{h}_1 \sim_k \mathbf{h}_2$), a full-rank transition Jacobian ensures that MTP induces a stable contractive force with $\dot{\mathcal{D}} \leq 0$. The resulting geometric flow is governed by $\mathbf{KS}$, where $\mathbf{K}$ is the NTK and $\mathbf{S}$ the pull-back Hessian. Although $\mathbf{KS}$ is generally non-symmetric, it is similar to a symmetric PSD matrix, implying real, non-negative eigenvalues and thus local convergence to a unified belief state.*

Intuition: This predictive coupling manifests as a geometric force that pulls together the representations of different histories whenever they lead to a common future.

These results demonstrate that MTP induces geometric contraction among representations sharing future dynamics. This effect facilitates the alignment of future-equivalent states (Section 5.2.1) and the compression of diverse histories into unified belief representations (Section 5.2.2). However, the contraction is inherently *outcome-driven*, ignoring the physical validity of intermediate transitions. As shown in our linear model (Section 5.1), MTP can induce transition weights toward unobserved states that happen to lead to the same target.

This phenomenon leads to *structural hallucinations*, where probability mass is incorrectly assigned to illegal shortcuts in latent space, causing the model to deviate from the true trajectory (Section 5.2.3). This theoretical gap motivates the development of our **LSE-MTP** framework (Section 4), which effectively anchors MTP-induced contraction to ground-truth latent trajectories.

4 What is LSE-MTP

We introduce **Latent Semantic Enhancement (LSE)**, a training framework built on Multi-Token Prediction (MTP) with a prediction horizon of K. Given backbone hidden states $\mathbf{h}_n \in \mathbb{R}^d$, we employ horizon-specific transition layers $\{\mathcal{T}_\phi^{(k-1)}\}_{k=2}^K$ to produce k-step predictive representations $\hat{\mathbf{h}}_{n,k} = \mathcal{T}_\phi^{(k-1)}(\mathbf{h}_n)$, with $\hat{\mathbf{h}}_{n,1} = \mathbf{h}_n$.

The training objective consists of three components. First, we apply a multi-step cross-entropy loss:

$$\mathcal{L}_{ce} = \sum_{k=1}^{K} \mathbb{E}_n \big[-\log P(u_{n+k} \mid \hat{\mathbf{h}}_{n,k}) \big]. \quad (4)$$

Second, a **latent consistency loss** aligns predictive representations with future backbone states:

$$\mathcal{L}_{latent} = \sum_{k=2}^{K} \mathbb{E}_n \big\| \hat{\mathbf{h}}_{n,k} - \mathbf{h}_{n+k-1} \big\|_2^2. \quad (5)$$

Third, a **semantic anchoring loss** aligns predictive representations with the target token embeddings $\mathbf{E}(\cdot)$:

$$\mathcal{L}_{semantic} = \sum_{k=2}^{K} \mathbb{E}_n \big\| \hat{\mathbf{h}}_{n,k} - \mathrm{sg}\big(\mathbf{E}(u_{n+k})\big) \big\|_2^2, \quad (6)$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator, and $\mathbf{E}(\cdot)$ denotes the model's embedding layer.

The full training objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_l \mathcal{L}_{latent} + \lambda_s \mathcal{L}_{semantic}. \quad (7)$$

Unless otherwise specified, we set $\lambda_l = \lambda_s = 0.1$.

At inference time, all transition layers and auxiliary losses are discarded, and decoding follows standard autoregressive NTP. The complete architecture of the model is illustrated in Figure 1.

5 Understanding Multi-Token Prediction in Modeling

In this section, we present two progressive experiments to empirically examine the theoretical analysis of Multi-Token Prediction (MTP) developed in Section 3.

5.1 How Multi-Token Prediction Induces Gradient Coupling

To isolate the gradient coupling mechanism of MTP from nonlinear confounders, we construct a minimal linear model. The states $\{A, B, C, D, E\}$ are represented as orthogonal basis vectors in $\mathbb{R}^5$, enabling a transparent analysis of how multi-step supervision reshapes local transition structure. The model has two learnable parameters: a backbone matrix $\boldsymbol{W}^B$ for one-step prediction ($\mathbf{h}_{t+1} = \boldsymbol{W}^B \mathbf{h}_t$) and, in the 2TP setting, an additional transition matrix $\boldsymbol{W}^T$ for predicting the state two steps ahead ($\mathbf{h}_{t+2} = \boldsymbol{W}^T \mathbf{h}_{t+1}$).

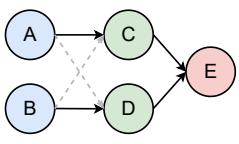

Figure 2: Two independent paths ($A \to C \to E$ and $B \to D \to E$) converging at a shared future E.

Table 1: **Representation alignment on ER and USG graphs.** Sim(F) and Gain denote cosine similarity and structure gain (Sim(F) - random baseline) for k-step future equivalent states.

Model	ER (Erdős–Rényi Graph)						USG (Urban Street Graph)					
	$k=2$		$k=3$		$k=4$		$k=2$		$k=3$		$k=4$	
	Sim(F)	Gain	Sim(F)	Gain	Sim(F)	Gain	Sim(F)	Gain	Sim(F)	Gain	Sim(F)	Gain
1TP	0.051	0.027	0.054	0.022	0.078	0.036	0.055	-0.005	0.082	0.018	0.072	0.005
2TP	0.232	**0.210**	0.102	0.074	0.094	0.062	0.264	**0.214**	0.126	0.066	0.112	0.048
3TP	0.229	0.195	0.194	**0.167**	0.136	0.107	0.249	0.197	0.244	**0.186**	0.148	0.083
4TP	0.223	0.176	0.201	0.162	0.204	**0.171**	0.230	0.178	0.235	0.180	0.222	**0.163**

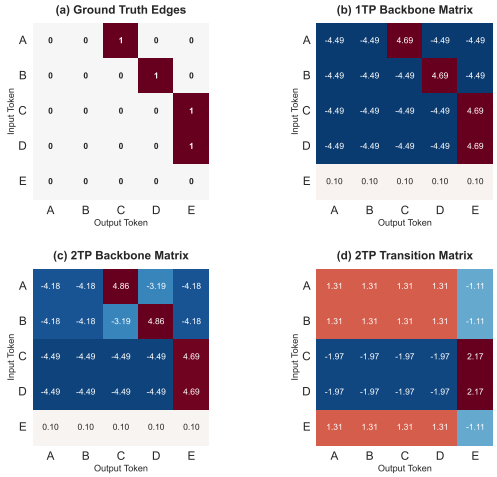

Figure 3: **Visualization of learned weights.** Under 2TP, unobserved cross-path transitions ($A \to D$, $B \to C$) are strengthened relative to 1TP.

The task contains two trajectories, $A \to C \to E$ and $B \to D \to E$ (Figure 2). We compare one-token prediction (1TP), optimizing only $\boldsymbol{W}^B$, with two-token prediction (2TP), jointly optimizing $\boldsymbol{W}^B$ and $\boldsymbol{W}^T$, with uniform initialization.

As shown in Figure 3, under 1TP, $\boldsymbol{W}^B$ learns only observed transitions like $A \to C$ (Figure 3b). Under 2TP, $\boldsymbol{W}^T$ captures two-step mappings from C, D to E (Figure 3d), while $\boldsymbol{W}^B$ also strengthens the unobserved transition $A \to D$ (Figure 3c).

This directly illustrates Theorem 2: since both C and D lead to the shared future target E, the gradient for predicting E backpropagates through $\boldsymbol{W}^T$ to both states, simultaneously strengthening the weights from A. Thus, when future targets coincide, MTP couples gradients across paths and updates transitions absent from the training data.

5.2 Representation Alignment under Multi-Token Supervision

We next investigate how multi-step supervision affects hidden state alignment.

Representation alignment is evaluated on two types of graphs:

- **ER (Erdős–Rényi Graphs)**: Random directed graphs capturing pure topological structure without spatial semantics.
- **USG (Urban Street Graphs)**: Planar road networks with node IDs reflecting approximate geography, enabling assessment of both topological and spatial continuity (Barthelemy and Boeing, 2025).

The navigation task is framed as conditional sequence generation. Given a start node S and a goal node G, forming the context $[S, G]$, the model autoregressively predicts stepwise increments $(\text{inc}_1, \ldots, \text{inc}_T)$, where each increment represents an action and node IDs are computed recursively as $u_t = u_{t-1} + \text{inc}_t$. A trajectory is valid if each increment corresponds to an existing edge (u_{t-1}, u_t) and the final node reaches the goal $u_T = G$. During training, sequences $[S, G, \text{inc}_1, \ldots, \text{inc}_T]$ serve as both input and autoregressive targets.

Reachable node pairs are split into 90% training and 10% test sets, with training paths generated via K-shortest paths, detours, and corrective strategies. On 100-node graphs, a 6-layer Transformer (6 attention heads, hidden dimension 120) is trained for 20,000 iterations, achieving ∼97% accuracy. This confirms that the model sufficiently masters the task to support subsequent representation analysis. The code is available at `https://github.com/QiminZhong/LSE-MTP`.

5.2.1 States with the Same Future Become Aligned

To quantify the effect of multi-step supervision on representation alignment, we introduce **Structure Gain**, measuring how closely states leading to the same future are embedded in latent space. The metric focuses on *k-step future equivalent* state pairs

Table 2: **Belief compression on ER and USG graphs.** Values report hidden-state similarity for trajectories sharing the same goal G and next-step position P under different control conditions; $=$ denotes "same" and $\neq$ denotes "different".

Model	K	ER (Erdős–Rényi Graph)				USG (Urban Street Graph)			
		$G=, P=$	$G=, P\neq$	$G\neq, P=$	Baseline	$G=, P=$	$G=, P\neq$	$G\neq, P=$	Baseline
NTP (1TP)	1	0.29	0.11	0.09	0.01	0.22	0.09	0.10	0.03
MTP	2	0.39	0.23	0.11	0.05	0.28	0.10	0.11	0.02
MTP	3	0.43	0.28	0.14	0.07	0.30	0.11	0.10	0.02
MTP	4	**0.44**	**0.30**	**0.15**	**0.08**	**0.32**	**0.12**	**0.09**	**0.03**
LSE-MTP	2	0.40	0.25	0.12	0.05	0.34	0.13	0.16	0.05
LSE-MTP	3	0.44	0.31	0.13	0.06	0.37	0.14	0.17	0.06
LSE-MTP	4	**0.46**	**0.34**	**0.16**	**0.09**	**0.38**	**0.15**	**0.17**	**0.06**

($\mathbf{h}_1 \sim_k \mathbf{h}_2$)—corresponding to the same token at step k but with different next-step targets—thus removing the confounding effect of immediate target agreement. Structure Gain is defined as the improvement in average cosine similarity of such state pairs relative to a random baseline. We compare a standard next-token prediction model (NTP, 1TP) with multi-token prediction models (MTP) trained with different prediction horizons $K \in \{2, 3, 4\}$, and evaluate at $k \in \{2, 3, 4\}$.

In each experiment, we randomly sample 4,000 training trajectories, extract normalized hidden states from the final Transformer layer, and construct pairs satisfying k-step future equivalence.

Table 1 shows that NTP exhibits low structure gain, indicating poor alignment of states sharing the same future. MTP models achieve substantially higher structure gain, with the effect strongest when training and evaluation horizons match (k). This trend supports Lemma 1: multi-token prediction induces cross-path gradient coupling, progressively converging states that lead to the same future in latent space.

5.2.2 Path Histories Are Compressed into a Unified Belief Representation

Beyond aligning future-equivalent states, we further investigate whether the model compresses diverse path histories into a unified internal representation. To this end, we introduce the **Belief Compression** metric, which quantifies the similarity of hidden states corresponding to trajectories that share the same goal G and, at the next step, reach the same position P, avoiding bias from identical immediate actions. This metric assesses whether the model can abstract away variations from different traversal histories and form a coherent internal *belief state*.

In our experiments, we randomly sample 4,000 training paths and evaluate all models on the same dataset. To examine the influence of goal and positional information in the representations, we introduce three control groups: (a) same goal, different positions ($G=, P\neq$); (b) different goals, same position ($G\neq, P=$); (c) different goals, different positions (baseline).

Table 2 summarizes the results. As the prediction horizon increases, MTP models exhibit higher hidden-state similarity under the same-goal, same-position condition ($G=, P=$), indicating that diverse path histories are compressed into a consistent belief representation. In contrast, the control settings and baseline show only minor increases, suggesting that compression is primarily driven by shared future outcomes.

5.2.3 A Pitfall: Probability Coupling in Next-Step Predictions

While MTP promotes representational alignment, it can introduce a teleological bias where the model prioritizes future outcomes over immediate constraints. Theorem 2 indicates that when distinct action sequences converge on the same future action token f, MTP induces predictive coupling within the next-step distribution. This effect can blur the distinction between feasible increments and illegal shortcuts—action tokens that move toward f but are invalid at the current state.

We evaluate this behavior using $10,000$ samples. Each test case involves a pair of action tokens (a, a') that share a common future action token f within two to four steps. In each pair, a is a valid increment along a legal edge, while a' is an illegal shortcut to an unconnected node. The model's performance is measured by **Illegal Shortcut Probability (ISP)**, the probability of the forbidden in-

Table 3: **Next-step probability coupling on ER and USG graphs.** ISP and Legal Prob report the probability of illegal shortcuts and valid actions for trajectories sharing a common future.

Model	K	ER (Erdős–Rényi Graph)		USG (Urban Street Graph)	
		ISP ↓	Legal Prob ↑	ISP ↓	Legal Prob ↑
NTP (1TP)	1	2.7×10^{-5}	**0.995**	2.2×10^{-5}	**0.998**
MTP	2	4.2×10^{-5}	0.994	4.9×10^{-5}	0.996
MTP	3	7.8×10^{-5}	0.992	7.3×10^{-5}	0.994
MTP	4	$\mathbf{1.04 \times 10^{-4}}$	0.985	$\mathbf{1.33 \times 10^{-4}}$	0.989
LSE-MTP	2	3.0×10^{-5}	0.995	4.1×10^{-5}	0.997
LSE-MTP	3	5.1×10^{-5}	0.993	4.8×10^{-5}	0.996
LSE-MTP	4	$\mathbf{6.3 \times 10^{-5}}$	0.990	$\mathbf{8.2 \times 10^{-5}}$	0.994

crement a', and **Legal Prob**, the total probability assigned to all valid actions.

Even a single token prediction error can cause the entire sequence to fail. As shown in Table 3, ISP gradually increases while Legal Prob decreases as the prediction horizon grows. The ISP reported in the table counts only illegal actions pointing to a single future token f, but the overall decline in Legal Prob reflects the cumulative effect of probability coupling across all potential illegal actions.

This observation is consistent with our theoretical perspective: while MTP promotes alignment of trajectories sharing a common future, it also blurs distinctions among feasible next-step predictions, resulting in illegal shortcuts due to prioritizing future alignment over immediate constraints.

Remark. Although the above experiments only present results from a single ER graph and a single USG graph, these phenomena are consistently observed across all generated graph instances.

6 Why LSE-MTP

The core motivation for LSE-MTP is to mitigate the *teleological bias* inherent in standard Multi-Token Prediction (MTP) under discrete-token supervision. In standard MTP, the gradient from the cross-entropy loss $\mathcal{L}_{ce}$ is focused solely on the discrete target token u_{n+k}, creating a "blind spot" regarding the feasibility of the intermediate path. This often encourages the model to adopt illegal shortcuts in latent space that violate structural constraints of the environment.

LSE-MTP addresses this issue by using the future hidden state $\mathbf{h}_{n+k-1}$ as a topological anchor. Since both $\hat{\mathbf{h}}_{n,k}$ and $\mathbf{h}_{n+k-1}$ are decoded by the shared output head to predict the same future token u_{n+k}, they are encouraged to occupy a consistent position in latent space. Targeting $\mathbf{h}_{n+k-1}$ is advantageous because it is generated through teacher forcing, thereby incorporating the ground-truth tokens $u_{n+1:n+k-1}$ along the path. This idea draws inspiration from Goyal et al. (2016), where teacher-forced hidden states serve as a continuous supervisor to regularize the model's self-generated trajectories. By aligning $\hat{\mathbf{h}}_{n,k}$ to $\mathbf{h}_{n+k-1}$, the latter acts as a grounded proxy that captures the structural rules of the environment that a "jump-step" prediction might otherwise bypass. In practice, LSE-MTP incurs almost zero additional computational cost compared to standard MTP, as detailed in Appendix D.

This alignment mechanism also resonates with the principles of the Joint-Embedding Predictive Architecture (JEPA) (LeCun and Courant, 2022) and Contrastive Predictive Coding (CPC) (van den Oord et al., 2018), which advocate predicting future dynamics in latent space rather than in the observation space. By performing latent backpropagation, structural information from the true trajectory is directly injected into the predictive transition layers. To stabilize training, the semantic loss $\mathcal{L}_{semantic}$ acts as a complementary regularizer, anchoring predictions to the static embedding manifold. This dual-grounding mechanism mitigates hallucinations by enhancing **Belief Compression** for identical states while simultaneously reducing **Illegal Shortcut Probability (ISP)** and increasing the probability assigned to valid actions, as evidenced in Tables 2 and 3. A more comprehensive sensitivity analysis of hyperparameters and the generalizability of LSE-MTP to unseen paths can be found in Appendix E.

Table 4: **Evaluation results on real-world Manhattan Taxi Ride Modeling.** Values are reported as mean (standard deviation).

Model	Valid Trajectories	Current State Probe	State-wise Similarity	Compression Precision	Distinction Precision	Distinction Recall	Detour Robustness
Sample Size	*1000 trials*	*1000 seqs*	*5000 trials*	*1000 trials*	*1000 trials*	*1000 trials*	*1000 trials*
1TP (baseline)	0.993 (0.003)	0.926 (0.001)	0.693 (0.139)	0.108 (0.011)	**0.357** (0.015)	0.210 (0.010)	0.692 (0.016)
4TP	0.997 (0.002)	0.964 (0.000)	0.722 (0.119)	0.119 (0.011)	0.298 (0.014)	0.195 (0.010)	0.708 (0.014)
8TP	0.995 (0.002)	0.964 (0.000)	0.820 (0.098)	0.114 (0.011)	0.293 (0.014)	0.182 (0.009)	0.716 (0.014)
LSE-4TP	0.997 (0.002)	0.943 (0.001)	0.791 (0.127)	0.135 (0.012)	0.327 (0.015)	**0.213** (0.011)	0.727 (0.014)
LSE-8TP	**0.998** (0.001)	**0.967** (0.000)	**0.851** (0.091)	**0.143** (0.012)	0.285 (0.014)	0.201 (0.010)	**0.733** (0.014)
True world model	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

7 Evaluation on Real-World Manhattan Taxi Ride Modeling

We evaluate our model on the Manhattan taxi trajectory benchmark introduced by Vafa et al. (2024), where city streets are abstracted as a graph with explicit topological constraints. Given a start and a destination, models are required to generate complete routes that are graph-consistent. This benchmark is suited for assessing the coherence of latent world models, as it reveals failures that are hard to detect with next-step prediction alone, such as infeasible paths or broken connectivity.

We train and evaluate the model on the shortest-paths dataset derived from this benchmark. All models adopt a Transformer architecture with 12 layers, 12 attention heads, and 768-dimensional embeddings, and are trained for 30 epochs to ensure convergence.

Most of the following metrics are adopted from Vafa et al. (2024) to assess the model's world modeling capability. (1) **Valid Trajectories** measures the fraction of complete sequences generated on unseen start–goal pairs that satisfy all street topology constraints and successfully reach the destination. (2) **Current State Probe** evaluates the accuracy of a linear classifier trained to predict the current node from the final-layer hidden representation. (3) **State-wise Similarity** computes the average cosine similarity between final-layer hidden states when two different paths reach the same node with the same goal. (4) **Compression Precision** is the fraction of continuations generated from one path that are assigned a prediction probability above a threshold ($\epsilon = 0.01$) under the other path's context, when two paths reach the same node with the same goal. (5) **Distinction Precision** measures, for two paths that differ in node or goal, the fraction of continuations that receive probability above $\epsilon = 0.01$ for only one path and correctly reflect the underlying map legality. (6) **Distinction Recall** evaluates, for continuations that are legal for only one of the two paths in the true map, the proportion of cases where the model correctly assigns a probability above $\epsilon = 0.01$ to one path and below the threshold to the other. Finally, (7) **Detour Robustness** computes the fraction of generated trajectories that remain valid and reach the goal when random non-Top-1 but legal turns are injected during generation with fixed probabilities $p = 0.01$.

Table 4 presents the evaluation on real-world Manhattan taxi trajectories. Multi-Token Prediction (MTP) improves both **state-wise similarity** and **compression precision**, indicating that trajectories sharing future dynamics are mapped to more consistent latent representations. This demonstrates that MTP effectively captures shared-future structure in the latent space. However, this increased alignment comes with a slight decrease in **distinction precision**, reflecting the inherent trade-off between aligning shared-future trajectories and preserving fine-grained state differences.

Incorporating LSE as a constraint on MTP mitigates this trade-off by grounding latent states in teacher-forced future representations. LSE further enhances **compression precision** while preserving or even boosting **distinction precision**, yielding a more balanced latent space that aligns shared-future trajectories without collapsing structurally relevant distinctions. The improved **detour robustness** also indicates that the learned latent dynamics are coherent and resilient to trajectory perturbations, enabling more robust trajectory planning.

8 Conclusion and Discussion

In this work, we study how multi-token prediction (MTP) shapes the internal representations of sequence models for latent world modeling. Our theoretical and empirical analyses reveal a key ten-

sion: while MTP promotes convergence toward shared-future belief states, discrete token supervision can induce structural hallucinations that disrupt latent dynamics. To address this, we propose Latent Semantic Enhancement MTP (LSE-MTP), which grounds multi-step predictions in teacher-forced latent trajectories and semantic embeddings. Experiments on synthetic graphs and real-world Manhattan taxi data show that LSE-MTP improves representation alignment, belief compression, and robustness, while reducing illegal shortcuts.

These results underscore that token-level accuracy alone is insufficient for coherent world modeling. By enforcing structurally consistent latent trajectories, LSE-MTP effectively bridges discrete supervision and continuous representations. This enables models to extend their predictive horizon while better preserving the local constraints that define the environment.

These structural challenges also apply to large-scale NLP tasks. In open-ended language, environmental constraints are not explicitly defined but emerge from an implicit logical and semantic manifold that governs coherence, causality, and plausibility. By leveraging teacher-forced hidden states, LSE-MTP captures coherent semantic trajectories along this manifold, modeling stepwise dependencies and gradual contextual evolution. Anchoring multi-step latent predictions to these trajectories, LSE-MTP provides a structural alignment signal that mitigates abrupt semantic shifts, encouraging the model to better integrate intermediate contextual cues and improve long-horizon coherence.

Such latent-space regularization is particularly crucial for tasks that require precise state tracking, such as narrative understanding, code generation, or mathematical reasoning (Kim and Schuster, 2023; Li et al., 2025). These tasks demand that models maintain consistent representations of entities, variables, or arguments over extended contexts. By regularizing latent trajectories, LSE-MTP helps transform language models from local pattern matchers into coherent internal simulators capable of reliable long-horizon reasoning.

9 Limitations

First, our experimental evaluation is primarily focused on structured graph navigation and path-planning tasks, with its applicability to open-ended natural language problems with higher levels of abstraction and more complex semantic dynamics not yet fully explored. Second, we have only analyzed and experimented with the widely used MTP model, without conducting a systematic comparison with other models and methods aimed at enhancing latent representation consistency, such as reinforcement learning objectives, contrastive representation learning, or explicit state-space modeling. Finally, our theoretical perspective relies on a linearized gradient flow approximation, which, while capturing the core trends of the training dynamics, may not fully reflect the complex nonlinear behavior of large-scale Transformer models.

Acknowledgments

This work is supported by thet National Natural Science Foundation of China (Grant No. 62276171, 62476173, 62532007), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011938 and 2020B1515120028), Shenzhen Fundamental Research Project (Grant No. ZDCY20250901110940006, JCYJ20240813-141503005, JCYJ20240813142610014) Major Special Project for Philosophy and Social Sciences Research of the Ministry of Education (Grant No. 2025JZDZ010). CCF-Huawei Populus Grove Fund (Grant No. CCF-HuaweiFM2024004). Mingyang Zhou and Hao Liao are the corresponding authors.

References

Gregor Bachmann and Vaishnavh Nagarajan. 2024. The pitfalls of next-token prediction. In *Forty-first International Conference on Machine Learning*.

Marc Barthelemy and Geoff Boeing. 2025. Universal model of urban street networks. *Physical Review Letters*, 135:137401.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple large language model inference acceleration framework with multiple decoding heads. In *Forty-first International Conference on Machine Learning*.

Lang Cao. 2024. GraphReason: Enhancing reasoning capabilities of large language models through a graph-based verification approach. In *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pages 1–12, Bangkok, Thailand. Association for Computational Linguistics.

Lénaïc Chizat, Edouard Oyallon, and Francis R. Bach. 2019. On lazy training in differentiable programming. In *Conference on Neural Information Processing Systems*, pages 2933–2943.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a graph: Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations*.

Arvid Frydenlund. 2025. Language models, graph searching, and supervision adulteration: When more supervision is less and how to make more more. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29011–29059, Vienna, Austria. Association for Computational Linguistics.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. 2024. Learning and Leveraging World Models in Visual Representation Learning. *arXiv e-prints*, arXiv:2403.00504.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. 2024. Better & faster large language models via multi-token prediction. In *Forty-first International Conference on Machine Learning*.

Anirudh Goyal, Alex Lamb, Ying Zhang, Saizheng Zhang, Aaron C. Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Prakhar Gupta, Henry Conklin, Sarah-Jane Leslie, and Andrew Lee. 2025. Better World Models Can Lead to Better Post-Training Performance. *arXiv e-prints*, arXiv:2512.03400.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

David Ha and Jürgen Schmidhuber. 2018. World models. *CoRR*, abs/1803.10122.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2025. Mastering diverse control tasks through world models. *Nature*, 640:647–653.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv e-prints*, arXiv:1503.02531.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Yann LeCun and Courant. 2022. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. In *Proceedings of the International Conference on Machine Intelligence*.

Anson Lei, Bernhard Schölkopf, and Ingmar Posner. 2023. Variational causal dynamics: Discovering modular world models from interventions. *Transactions on Machine Learning Research*.

Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. 2025. (how) do language models track state? In *Forty-second International Conference on Machine Learning*.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. 2023b. Can language models understand physical concepts? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11843–11861, Singapore. Association for Computational Linguistics.

Michael L. Littman and Richard S. Sutton. 2001. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context Learning and Induction Heads. *arXiv e-prints*, arXiv:2209.11895.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Jeffrey Pennington and Pratik Worah. 2017. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv e-prints*, arXiv:2112.11446.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Fourth International Conference on Learning Representations (ICLR 2016), Conference Track Proceedings, San Juan, Puerto Rico*.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv e-prints*, arXiv:1312.6120.

Jürgen Schmidhuber. 1990. Making the world differentiable: on using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments. *Forschungsberichte, TU Munich*, FKI 126 90:1–26.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, L. Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. 2019. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604–609.

David Silver, Hado van Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David P. Reichert, Neil C. Rabinowitz, André Barreto, and Thomas Degris. 2017. The predictron: End-to-end learning and planning. In *Thirty-fourth International Conference on Machine Learning (ICML 2017), Sydney, NSW, Australia*, volume 70 of *Proceedings of Machine Learning Research*, pages 3191–3199. PMLR.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2025. On the self-verification limitations of large language models on reasoning and planning tasks. In *The Thirteenth International Conference on Learning Representations*.

Jayden Teoh, Manan Tomar, Kwangjun Ahn, Edward S. Hu, Pratyusha Sharma, Riashat Islam, Alex Lamb, and John Langford. 2025. Next-Latent Prediction Transformers Learn Compact World Models. *arXiv e-prints*, arXiv:2511.05963.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. 2024. Evaluating the world model implicit in a generative model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv e-prints*, arXiv:1807.03748.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388, Bangkok, Thailand. Association for Computational Linguistics.

Xixi Wu, Yifei Shen, Caihua Shan, Kaitao Song, Siwei Wang, Bohang Zhang, Jiarui Feng, Hong Cheng, Wei Chen, Yun Xiong, and Dongsheng Li. 2024. Can graph learning improve planning in large language model-based agents? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Charlie Wyatt, Aditya Joshi, and Flora Salim. 2025. Alternatives To Next Token Prediction In Text Generation – A Survey. *arXiv e-prints*, arXiv:2509.24435.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On Layer Normalization in the Transformer Architecture. *arXiv e-prints*, arXiv:2002.04745.

Yongjing Yin, Junran Ding, Kai Song, and Yue Zhang. 2024. Semformer: Transformer language models with semantic planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18669–18680, Miami, Florida, USA. Association for Computational Linguistics.

Guangyao Zhai, Xingyuan Zhang, and Nassir Navab. 2025. Recurrent world model with tokenized latent states. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Fu Zhang, Jinghao Lin, and Jingwei Cheng. 2024. SALMON: A structure-aware language model with logicality and densification strategy for temporal knowledge graph reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8761–8774, Miami, Florida, USA. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

Qimin Zhong, Hao Liao, Siwei Wang, Mingyang Zhou, Xiaoqun Wu, Rui Mao, and Wei Chen. 2025. Understanding and Enhancing the Planning Capability of Language Models via Multi-Token Prediction. *arXiv e-prints*, arXiv:2509.23186.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. In *Forty-first International Conference on Machine Learning*.

Appendix

A Related Works

A.1 World Models in Language Modeling

The debate over whether Large Language Models (LLMs) are "stochastic parrots" (Bender et al., 2021) or possess emergent *world models* remains central to NLP (Li et al., 2023a; Patel and Pavlick, 2022). Probing studies suggest that neural language models can indeed develop implicit representations of meaning and world states even when trained solely on text (Li et al., 2021). While Transformers can internalize structural invariants like game states (Li et al., 2023a) or geographical coordinates (Gurnee and Tegmark, 2024), and even demonstrate a nascent grasp of fundamental physical concepts (Li et al., 2023b), they are often reactive rather than truly predictive. Recent studies highlight failures in multi-step causal reasoning and state tracking (Valmeekam et al., 2023; Wu et al., 2024), showing fragility under structural perturbations (Vafa et al., 2024). Beyond simple internalization, discovering structured and modular world models from data is a prerequisite for reliable planning. This shift toward causal modularity facilitates the emergence of consistent belief states (Lei et al., 2023), encouraging paradigms that frame reasoning as an explicit planning process over an internal world model (Hao et al., 2023), and motivating architectures that move from local co-occurrence statistics toward explicit environment dynamics (Hafner et al., 2025; LeCun and Courant, 2022).

A.2 Multi-Token Prediction

Multi-Token Prediction (MTP) improves upon standard next-token prediction by supervising multiple future tokens, which enhances performance on reasoning benchmarks (Gloeckle et al., 2024). By incentivizing the model to anticipate future sequence fragments during pre-training, MTP builds on the intuition that future n-gram prediction can foster more robust contextual representations and sequence-level planning (Qi et al., 2020). Theoretically, MTP fosters longer-range dependencies and "look-ahead" foresight (Olsson et al., 2022; Cai et al., 2024). In graph-based planning, MTP prevents local shortcuts via multi-hop trajectory mixtures (Frydenlund, 2025) and captures transitive reachability by encoding multi-step adjacency (Zhong et al., 2025). This approach is rooted in earlier sequence-level optimization techniques like MIXER (Ranzato et al., 2016), which aim to mitigate exposure bias (Bachmann and Nagarajan, 2024) by narrowing the distributional gap between teacher-forcing training and autoregressive inference (Zhang et al., 2019), preventing the accumulation of errors during rollout. However, behavioral gains do not guarantee latent consistency, as models may still learn shortcuts that bypass underlying environmental rules (Geirhos et al., 2020). Our work explores this risk of "structural hallucinations" and proposes latent grounding as a necessary stabilizer.

A.3 Latent Consistency and State-Space Alignment

Reliable world modeling requires internal representations to evolve consistently with environmental dynamics. In autonomous intelligence, architectures such as JEPA (LeCun and Courant, 2022) and Dreamer (Hafner et al., 2025) advocate predicting in latent space rather than observation space, thereby filtering out task-irrelevant noise. NextLat (Teoh et al., 2025) extends this principle to Transformers through self-supervised latent-state prediction, encouraging models to learn compressed belief states and form compact internal world models. Recent efforts have further explored aligning hidden states with symbolic world structures (Zhu et al., 2024; Garrido et al., 2024) or leveraging tokenized latent states (Zhai et al., 2025). Similarly, methods such as Semformer (Yin et al., 2024) promote semantic planning by predicting future latent representations generated by an auxiliary autoencoder. Building on these insights, LSE-MTP draws inspiration from knowledge distillation (Hinton et al., 2015) and consistency models (Song et al., 2023) to improve transition consistency by anchoring predictions to ground-truth hidden states.

A.4 Graph-based Reasoning and Navigation

Graphs provide a rigorous testbed for world models due to their explicit transition rules (Li et al., 2023a; Wu et al., 2024). Navigating these environments requires models to maintain logical consistency through structure-aware architectures that can handle complex relational constraints and densification (Zhang et al., 2024). Recent benchmarks, such as Manhattan taxi trajectories (Vafa et al., 2024), require models to adhere to real-world topology over long horizons. Despite generating fluent paths, LLMs often prioritize statistical patterns over topological constraints, leading to planning failures dur-

ing detours (Valmeekam et al., 2023; Stechly et al., 2025). Such failures emphasize the need for graph-based verification mechanisms that can audit reasoning chains against the underlying connectivity to ensure path validity (Cao, 2024). While specialized fine-tuning or prompting techniques like structuring internal evidence into a graph of thoughts can improve performance (Fatemi et al., 2024; Wen et al., 2024), the fundamental challenge of latent state legality remains. We utilize synthetic and real-world graphs to demonstrate how latent grounding prevents models from taking illegal shortcuts that violate connectivity.

B Derivations and Proofs

This appendix rigorously characterizes the gradient dynamics, detailing assumptions, validity conditions, and proofs for the established theorems.

B.1 Validity of Linearized Analysis

To understand how the training process shapes internal representations, we analyze the model's behavior through the lens of the linearized regime, also known as *lazy training* (Chizat et al., 2019). Deep neural networks, like Transformers, are notoriously complex and non-linear, making their training dynamics difficult to track mathematically. However, a key theoretical insight in deep learning is that as a network becomes sufficiently wide, its individual weights θ only need to change by a tiny amount from their initial values θ_0 to significantly reduce the training loss. In this "lazy" state, we can accurately approximate the network's output, specifically the hidden state $f_\theta(H)$, using a first-order Taylor expansion:

$$f_\theta(H) \approx f_{\theta_0}(H) + \nabla_\theta f_{\theta_0}(H)^\top (\theta - \theta_0). \quad (8)$$

This approximation effectively treats the complex network as a linear model during the early stages of training. The primary advantage of this approach is that it allows us to define the **Neural Tangent Kernel (NTK)**, a mathematical object that remains approximately constant during training. The NTK acts like a *geometric map* of the representation space, determining how an update on one input, such as a specific history H_i, influences the representation of another, H_j. By assuming this kernel is stable, we can derive closed-form proofs for how gradients flow through the model.

While real-world, finite-width Transformers eventually move beyond this linear phase to perform *feature learning*, the linearized analysis remains a powerful tool for our purposes. It provides a clear, qualitative explanation of the instantaneous directional pressure, which represents the immediate "force" that the Multi-Token Prediction (MTP) objective exerts on hidden states. By capturing the direction in which the loss function pushes representations at any given moment, this framework reveals the mathematical root of the gradient coupling and representational contraction observed in our empirical experiments.

B.2 Evolution Dynamics and Notation

To track how hidden states $\mathbf{h}$ evolve during training, we study their dynamics under **gradient flow**. Let $\mathbf{h} = f_\theta(H) \in \mathbb{R}^d$ denote the hidden representation of a history H. The continuous-time optimization of parameters is controlled by a learning rate $\eta > 0$, and the weight evolution is given by:

$$\dot{\theta} = \frac{d\theta}{dt} = -\eta \nabla_\theta \mathcal{L}, \quad (9)$$

where $\mathcal{L}$ is the loss function. Applying the chain rule, the velocity of the hidden state (the rate of change over time) satisfies:

$$\dot{\mathbf{h}} = \nabla_\theta f_\theta(H)\dot{\theta} = -\eta \nabla_\theta f_\theta(H) \nabla_\theta \mathcal{L}. \quad (10)$$

Since the loss depends on the weights θ primarily through the representation $\mathbf{h}$, we can further decompose the weight gradient using the chain rule again:

$$\nabla_\theta \mathcal{L} = \nabla_\theta f_\theta(H)^\top \nabla_{\mathbf{h}} \mathcal{L}. \quad (11)$$

Substituting this back into the velocity equation yields:

$$\begin{aligned} \dot{\mathbf{h}} &= -\eta \left[\nabla_\theta f_\theta(H) \nabla_\theta f_\theta(H)^\top \right] \nabla_{\mathbf{h}} \mathcal{L} \\ &= -\eta\, \mathbf{K}(\mathbf{h}, \mathbf{h}) \nabla_{\mathbf{h}} \mathcal{L}, \end{aligned} \quad (12)$$

where

$$\mathbf{K}(\mathbf{h}_i, \mathbf{h}_j) = \nabla_\theta f_\theta(H_i) \nabla_\theta f_\theta(H_j)^\top \in \mathbb{R}^{d \times d} \quad (13)$$

is the NTK matrix block, which measures the geometric correlation between the gradients of two different history samples H_i and H_j.

Intuitively, this expression shows that hidden-state updates are driven by the loss gradient $\nabla_{\mathbf{h}} \mathcal{L}$ and modulated by the kernel $\mathbf{K}$, which captures geometric correlations between different histories. When $\mathbf{K}$ exhibits strong cross-history coupling, the corresponding representations are forced to evolve jointly, providing the core mechanism behind the representational contraction.

B.3 Proof of Theorem 1

Theorem 1 *Under the NTP loss $\mathcal{L}_{NTP}$, the contractive condition $\dot{\mathcal{D}} \leq 0$ holds primarily for 1-step equivalent states ($\mathbf{h}_1 \sim_1 \mathbf{h}_2$). For states with different next-step targets, the gradients $\nabla_{\mathbf{h}}\mathcal{L}$ tend to point in opposite directions, preserving representational separation.*

Proof. To analyze the convergence of hidden states, we define the representational distance as $\mathcal{D} = \|\mathbf{h}_1 - \mathbf{h}_2\|^2$. The time derivative of this distance, which represents the rate at which states move toward or away from each other, is calculated as:

$$\dot{\mathcal{D}} = \frac{d}{dt}\|\mathbf{h}_1 - \mathbf{h}_2\|^2 = 2(\mathbf{h}_1 - \mathbf{h}_2)^\top(\dot{\mathbf{h}}_1 - \dot{\mathbf{h}}_2). \tag{14}$$

By substituting the hidden-state velocity formula $\dot{\mathbf{h}} = -\eta\mathbf{K}\nabla_{\mathbf{h}}\mathcal{L}$ derived in Section B.2, the dynamics are expressed as:

$$\dot{\mathcal{D}} = -2\eta(\mathbf{h}_1 - \mathbf{h}_2)^\top\big[\mathbf{K}(\mathbf{h}_1, \mathbf{h}_1)\nabla_{\mathbf{h}_1}\mathcal{L} - \mathbf{K}(\mathbf{h}_2, \mathbf{h}_2)\nabla_{\mathbf{h}_2}\mathcal{L}\big], \tag{15}$$

where η denotes the learning rate, $\mathcal{L}$ is the loss function, and $\mathbf{K}(\mathbf{h}_1, \mathbf{h}_1), \mathbf{K}(\mathbf{h}_2, \mathbf{h}_2)$ are the auto-kernel blocks for histories H_1, H_2.

Assumption 1 (Local Kernel Smoothness). For nearby states, we assume the kernel varies smoothly such that $\mathbf{K}(\mathbf{h}_1, \mathbf{h}_1) \approx \mathbf{K}(\mathbf{h}_2, \mathbf{h}_2) \approx \mathbf{K}$, where $\mathbf{K}$ is a positive semi-definite matrix. This assumption implies that the geometric properties of the representation space are locally stable, ensuring consistent sensitivity to parameter updates for nearby histories.

Using this assumption and letting $\Delta\mathbf{h} = \mathbf{h}_1 - \mathbf{h}_2$, the dynamics of the representational distance simplify to:

$$\dot{\mathcal{D}} \approx -2\eta\Delta\mathbf{h}^\top\mathbf{K}\left(\nabla_{\mathbf{h}_1}\mathcal{L} - \nabla_{\mathbf{h}_2}\mathcal{L}\right). \tag{16}$$

To further simplify the gradient difference term, we consider the gradient $\nabla_{\mathbf{h}}\mathcal{L}$ as a vector-valued function of $\mathbf{h}$. Since $\mathbf{h}_1$ and $\mathbf{h}_2$ are assumed to be in close proximity, we can apply a first-order Taylor expansion to the gradient $\nabla_{\mathbf{h}_1}\mathcal{L}$ around the point $\mathbf{h}_2$:

$$\nabla_{\mathbf{h}_1}\mathcal{L} \approx \nabla_{\mathbf{h}_2}\mathcal{L} + \left[\nabla^2_{\mathbf{h}}\mathcal{L}\right](\mathbf{h}_1 - \mathbf{h}_2), \tag{17}$$

where $\nabla^2_{\mathbf{h}}\mathcal{L}$ is the Hessian matrix of the loss function, denoted as $\mathbf{H}_{\text{loss}}$. This matrix captures the local curvature of the optimization landscape.

By rearranging the above expansion, we obtain an approximation for the gradient difference:

$$\nabla_{\mathbf{h}_1}\mathcal{L} - \nabla_{\mathbf{h}_2}\mathcal{L} \approx \mathbf{H}_{\text{loss}}\Delta\mathbf{h}. \tag{18}$$

Finally, substituting this approximation back into Eq. (16) yields the final quadratic form:

$$\dot{\mathcal{D}} \approx -2\eta\Delta\mathbf{h}^\top(\mathbf{K}\mathbf{H}_{\text{loss}})\Delta\mathbf{h}. \tag{19}$$

Conclusion for NTP. In Next-Token Prediction, the contractive condition $\dot{\mathcal{D}} \leq 0$ requires the gradients to converge toward a shared optimum. When target tokens differ ($u_{n+1} \neq u_{m+1}$), the gradients $\nabla_{\mathbf{h}_1}\mathcal{L}$ and $\nabla_{\mathbf{h}_2}\mathcal{L}$ point in opposite directions. As a result, $\Delta\mathbf{h}^\top(\nabla_{\mathbf{h}_1}\mathcal{L} - \nabla_{\mathbf{h}_2}\mathcal{L})$ becomes negative, leading to $\dot{\mathcal{D}} > 0$. Therefore, representations diverge unless they share the same target.

B.4 Proof of Theorem 2

Theorem 2 *Under the MTP loss $\mathcal{L}_{MTP}$, consider k-step future-equivalent states $\mathbf{h}_1 \sim_k \mathbf{h}_2$ with different immediate targets $u_{n+1} \neq u_{m+1}$. A k-step update on $\mathbf{h}_1$ induces a positive cross-update on the corresponding logit of $\mathbf{h}_2$, $\dot{z}_{y_1}(\mathbf{h}_2) > 0$, where the gradients $\nabla_{\mathbf{h}_1}\mathcal{L}_1^{(k)}$ and $\nabla_{\mathbf{h}_2}\mathcal{L}_1^{(k)}$ align through the cross-history NTK $\mathbf{K}(\mathbf{h}_1, \mathbf{h}_2)$, facilitating a predictive coupling that can partially blur the representational separation between distinct trajectories.*

Proof. To analyze predictive coupling in MTP, we track how the k-step loss $\mathcal{L}_1^{(k)}$ on history H_1 affects the logit $z_{y_1}(\mathbf{h}_2)$ for H_1's first future token y_1 in history H_2.

Under gradient flow, the parameter dynamics $\dot{\theta} = -\eta\nabla_\theta\mathcal{L}_1^{(k)}$ govern the evolution of $z_{y_1}(\mathbf{h}_2)$:

$$\begin{aligned}\frac{dz_{y_1}(\mathbf{h}_2)}{dt} &= \langle\nabla_\theta z_{y_1}(\mathbf{h}_2), \dot{\theta}\rangle \\ &= -\eta\langle\nabla_\theta z_{y_1}(\mathbf{h}_2), \nabla_\theta\mathcal{L}_1^{(k)}\rangle.\end{aligned} \tag{20}$$

Both gradients depend on θ only through the hidden representations $\mathbf{h}_2 = f_\theta(H_2)$ and $\mathbf{h}_1 = f_\theta(H_1)$:

$$\begin{aligned}\nabla_\theta z_{y_1}(\mathbf{h}_2) &= \nabla_{\mathbf{h}_2} z_{y_1}\,\nabla_\theta f_\theta(H_2), \\ \nabla_\theta\mathcal{L}_1^{(k)} &= \nabla_\theta f_\theta(H_1)^\top\nabla_{\mathbf{h}_1}\mathcal{L}_1^{(k)}.\end{aligned} \tag{21}$$

Substituting into Eq. (20) gives:

$$\frac{dz_{y_1}(\mathbf{h}_2)}{dt} = -\eta(\nabla_{\mathbf{h}_2} z_{y_1})^\top\mathbf{K}(\mathbf{h}_2, \mathbf{h}_1)\nabla_{\mathbf{h}_1}\mathcal{L}_1^{(k)}, \tag{22}$$

where

$$\mathbf{K}(\mathbf{h}_2, \mathbf{h}_1) = \nabla_\theta f_\theta(H_2) \nabla_\theta f_\theta(H_1)^\top \tag{23}$$

is the cross-history NTK block capturing geometric coupling between the hidden representations $\mathbf{h}_2$ and $\mathbf{h}_1$.

Assumption 2 (Structural Alignment). We assume that the transition layer Jacobian J_k preserves gradient orientation over k steps. Under this assumption, the k-step gradient from history H_1, $\nabla_{\mathbf{h}_1} \mathcal{L}_1^{(k)}$, lies in the subspace spanned by the cross-history NTK $\mathbf{K}(\mathbf{h}_2, \mathbf{h}_1)$ and the k-step gradient of H_2, $\nabla_{\mathbf{h}_2} \mathcal{L}_2^{(k)}$. This ensures predictable interactions between gradients from different histories in multi-token prediction.

Conclusion for MTP. The update direction is determined by the inner product in Eq. (20). Under the Structural Alignment Assumption, the gradient propagated from H_1 affects H_2 predictably. When H_1 and H_2 share the same k-step future token sequence, we can approximate

$$\nabla_{\mathbf{h}_2} z_{y_1} \approx -\alpha \nabla_{\mathbf{h}_2} \mathcal{L}_1^{(k)}, \quad \alpha > 0, \tag{24}$$

where the negative sign reflects that gradient descent on the loss $\mathcal{L}_1^{(k)}$ decreases the loss but increases the corresponding logits.

Substituting this into Eq. (22) then yields a positive cross-update:

$$\begin{aligned} \frac{dz_{y_1}(\mathbf{h}_2)}{dt} &= -\eta (\nabla_{\mathbf{h}_2} z_{y_1})^\top \mathbf{K}(\mathbf{h}_2, \mathbf{h}_1) \nabla_{\mathbf{h}_1} \mathcal{L}_1^{(k)} \\ &\approx -\eta \, (-\alpha \nabla_{\mathbf{h}_2} \mathcal{L}_1^{(k)})^\top \mathbf{K}(\mathbf{h}_2, \mathbf{h}_1) \nabla_{\mathbf{h}_1} \mathcal{L}_1^{(k)} \\ &= \eta \, \alpha \, (\nabla_{\mathbf{h}_2} \mathcal{L}_1^{(k)})^\top \mathbf{K}(\mathbf{h}_2, \mathbf{h}_1) \nabla_{\mathbf{h}_1} \mathcal{L}_1^{(k)} \\ &= \eta \, \alpha \, (\nabla_{\mathbf{h}_1} \mathcal{L}_1^{(k)})^\top \mathbf{K}(\mathbf{h}_1, \mathbf{h}_2) \nabla_{\mathbf{h}_2} \mathcal{L}_1^{(k)} > 0. \end{aligned} \tag{25}$$

Consequently, supervising H_1 via its k-step loss increases the probability that H_2 predicts the same sequence of future tokens, including y_1, even if H_2's own next-token target y_2 differs.

B.5 Proof of Lemma 1

Lemma 1 *For a pair of k-step future-equivalent states ($\mathbf{h}_1 \sim_k \mathbf{h}_2$), a full-rank transition Jacobian ensures that MTP induces a stable contractive force with $\dot{\mathcal{D}} \leq 0$. The resulting geometric flow is governed by $\mathbf{KS}$, where $\mathbf{K}$ is the NTK and $\mathbf{S}$ the pull-back Hessian. Although $\mathbf{KS}$ is generally non-symmetric, it is similar to a symmetric PSD matrix, implying real, non-negative eigenvalues and thus local convergence to a unified belief state.*

Proof. We examine the contractivity of the representational distance $\mathcal{D} = \|\mathbf{h}_1 - \mathbf{h}_2\|^2$ for histories that share a common k-step future. Let $\Delta\mathbf{h} = \mathbf{h}_1 - \mathbf{h}_2$ denote the difference between hidden representations.

Under gradient flow, the evolution of each hidden state is given by

$$\dot{\mathbf{h}} = -\eta \, \mathbf{K} \nabla_{\mathbf{h}} \mathcal{L}. \tag{26}$$

where $\mathbf{K}$ is the NTK capturing the sensitivity of hidden states to parameter updates.

The rate of change of the representational distance is obtained by differentiating $\mathcal{D} = \Delta\mathbf{h}^\top \Delta\mathbf{h}$ with respect to time:

$$\begin{aligned} \dot{\mathcal{D}} &= \frac{d}{dt}(\Delta\mathbf{h}^\top \Delta\mathbf{h}) \\ &= 2 \, \Delta\mathbf{h}^\top \frac{d}{dt}(\Delta\mathbf{h}) \\ &= 2 \, \Delta\mathbf{h}^\top (\dot{\mathbf{h}}_1 - \dot{\mathbf{h}}_2), \end{aligned} \tag{27}$$

where we used $\frac{d}{dt}(\mathbf{h}_1 - \mathbf{h}_2) = \dot{\mathbf{h}}_1 - \dot{\mathbf{h}}_2$.

Hence, the rate of change of the distance is determined by the projection of the hidden-state velocity difference onto the current difference $\Delta\mathbf{h}$, providing a direct measure of convergence or divergence between the two representations.

For multi-step prediction, the loss depends on the hidden states through k transition layers and a shared prediction head. Let $J_k = \frac{\partial \mathbf{z}}{\partial \mathbf{h}}$ denote the Jacobian from the hidden state $\mathbf{h}$ to the output logits $\mathbf{z}$, and $\mathbf{H}_{\text{head}} = \frac{\partial^2 \mathcal{L}}{\partial \mathbf{z}^2}$ be the Hessian of the prediction head.

Consider two hidden states $\mathbf{h}_1$ and $\mathbf{h}_2$ that are close in representation space. A first-order Taylor expansion of the gradient at $\mathbf{h}_1$ around $\mathbf{h}_2$ gives

$$\nabla_{\mathbf{h}_1} \mathcal{L} \approx \nabla_{\mathbf{h}_2} \mathcal{L} + \nabla_{\mathbf{h}}^2 \mathcal{L}\big|_{\mathbf{h}_2} (\mathbf{h}_1 - \mathbf{h}_2), \tag{28}$$

where $\nabla_{\mathbf{h}}^2 \mathcal{L}$ is the Hessian of the loss with respect to the hidden state. For multi-step prediction, the loss gradient is backpropagated through k transition layers, so the effective Hessian with respect to the original hidden state can be expressed via the chain rule as

$$\mathbf{S} = J_k^\top \mathbf{H}_{\text{head}} J_k, \tag{29}$$

where $\mathbf{H}_{\text{head}} = \nabla_{\mathbf{z}}^2 \mathcal{L}_{\text{head}}$ is the Hessian of the prediction head, and J_k maps perturbations in $\mathbf{h}$ to the output logits z through the $(k-1)$-th transition layer. Hence, the gradient difference between the two hidden states can be approximated as

$$\nabla_{\mathbf{h}_1} \mathcal{L} - \nabla_{\mathbf{h}_2} \mathcal{L} \approx \mathbf{S} \, (\mathbf{h}_1 - \mathbf{h}_2) = \mathbf{S} \, \Delta\mathbf{h}. \tag{30}$$

Intuitively, the pull-back Hessian $\mathbf{S}$ captures how local variations in the hidden state propagate through the transition layers and the prediction head to affect the loss, effectively defining a local metric for representational contraction.

Substituting the gradient approximation Eq. (30) into the distance dynamics Eq. (27) and using the gradient flow Eq. (26), we obtain

$$\dot{\mathcal{D}} \approx -2\eta\, \Delta\mathbf{h}^{\top}(\mathbf{KS})\, \Delta\mathbf{h}. \tag{31}$$

In this expression, $\mathbf{K}$ quantifies how changes in model parameters affect the hidden states, while $\mathbf{S}$ encodes how small variations in the hidden representations propagate through the transition layers and the prediction head to influence the loss. The product $\mathbf{KS}$ therefore defines a local metric that determines the rate and direction of contraction: the negative sign ensures that the component of the hidden-state difference $\Delta\mathbf{h}$ along sensitive directions is reduced over time. As a result, representations of histories that share a common future are drawn toward each other, forming a stable, contractive manifold in representation space.

A key theoretical concern is that, although both the kernel $\mathbf{K}$ and the pull-back Hessian $\mathbf{S}$ are symmetric positive semi-definite (PSD), their product $\mathbf{KS}$ is not guaranteed to be symmetric or PSD. To ensure that representations converge (i.e., $\dot{\mathcal{D}} \leq 0$), we must verify that all eigenvalues of $\mathbf{KS}$ are real and non-negative.

Assuming a locally strictly convex prediction head and a full-rank transition Jacobian, we have $\mathbf{S} \succ 0$. We can then perform a similarity transformation on $\mathbf{KS}$ using the square root of the Hessian:

$$\mathbf{S}^{1/2}(\mathbf{KS})\mathbf{S}^{-1/2} = \mathbf{S}^{1/2}\mathbf{K}\mathbf{S}^{1/2} \triangleq \tilde{\mathbf{K}}. \tag{32}$$

Since $\mathbf{K}$ is PSD and $\mathbf{S}^{1/2}$ is symmetric, $\tilde{\mathbf{K}}$ is symmetric and PSD. Because similar matrices share the same eigenvalues, we have

$$\lambda_i(\mathbf{KS}) = \lambda_i(\tilde{\mathbf{K}}) \geq 0. \tag{33}$$

This guarantees that the dynamical system has no unstable or divergent modes. The induced flow generates a stable contractive force in the metric defined by $\mathbf{S}$, attracting representations of histories sharing a common future toward each other. Hence, MTP establishes a stable manifold that formally proves representational contraction.

B.6 Boundary Conditions

The theoretical validity of the representational contraction depends on the numerical stability of the transition Jacobian J_k. If J_k were to suffer from rank-deficiency, the pull-back Hessian $\mathbf{S}$ would become singular, effectively halting the convergence of belief states. While this is a common failure mode in deep linear networks (Saxe et al., 2013; Pennington and Worah, 2017), modern Transformer architectures incorporate design elements that mitigate this risk.

Residual connections and LayerNorm collectively maintain a non-zero minimum singular value (Xiong et al., 2020), $\sigma_{\min}(J_k) > 0$, across the hidden layers. These architectural features ensure that $\mathbf{S}$ remains positive-definite, thereby preserving the gradient flow required for the emergence of consistent internal belief states. This indicates that our theoretical findings are well-supported by the structural properties of Transformer-based world models.

C Detailed Dataset Construction and Experimental Setup

This section describes the procedures for generating the graph environments and the trajectory datasets used in our experiments.

C.1 Graph Topology Construction

Two types of graphs are constructed: Erdős-Rényi (ER) random graphs and Planar Road Layout (USG) networks.

ER-Random Graphs. We generate directed graphs using $n = 100$ nodes and an edge probability $p = 0.04$. For directed acyclic graphs (DAGs), edges are permitted only according to a random topological ordering. The procedure is shown in Algorithm 1.

USG-Urban Street Graphs. These graphs are built using geometric triangulation and spatial relabeling. We set $n = 100$ and mesh density $\rho = 0.3$. The procedure is detailed in Algorithm 2.

C.2 Path Generation and Augmentation

We perform a 90/10 split on reachable node pairs for training and testing. The test set contains only unique reachable node pairs. For each training pair, we generate a diverse set of paths, including (i) shortest and top-K shortest paths, (ii) detour

Algorithm 1 ER-Random Graph Generation

Require: Nodes n, probability p, boolean is_dag
Ensure: Directed Graph $G = (V, E)$

$V \leftarrow \{0, \ldots, n-1\}$, $E \leftarrow \emptyset$
if is_dag **then**
  $\pi \leftarrow$ Random permutation of V
  $pos[v] \leftarrow$ index of v in $\pi, \forall v \in V$
  for each pair (u, v) where $u \neq v$ **do**
    if $pos[u] < pos[v]$ **and** rand$(0, 1) < p$ **then**
      $E \leftarrow E \cup \{(u, v)\}$
    end if
  end for
else
  for each pair (u, v) where $u \neq v$ **do**
    if rand$(0, 1) < p$ **then**
      $E \leftarrow E \cup \{(u, v)\}$
    end if
  end for
end ifreturn $G(V, E)$

Algorithm 2 USG-Urban Street Graph Generation

Require: Nodes n, density ρ
Ensure: Directed Graph G_{final}

$Pos \leftarrow \{(x_i, y_i) \mid x_i, y_i \sim \mathcal{U}(0, 1)\}_{i=0}^{n-1}$
$G_{base} \leftarrow$ DelaunayTriangulation(Pos)
$T_{mst} \leftarrow$ MinimumSpanningTree (G_{base}) ▷ Ensure connectivity
$E_{extra} \leftarrow G_{base}.edges \setminus T_{mst}.edges$
$E_{add} \leftarrow$ Sample $\lfloor |E_{extra}| \cdot \rho \rfloor$ edges from E_{extra}
$G_{sparse} \leftarrow (V, T_{mst}.edges \cup E_{add})$
$V_{sorted} \leftarrow$ Sort V by (x, y) lexicographically
$\phi(v) \leftarrow$ index of v in V_{sorted}
$G_{relabeled} \leftarrow$ Relabel G_{sparse} via mapping ϕ
$G_{final} \leftarrow$ Convert to Directed Graph
for each edge $\{u, v\} \in G_{relabeled}$ **do**
  Add (u, v) and (v, u) to G_{final}
end forreturn G_{final}

paths obtained by temporarily removing intermediate nodes, and (iii) recovery paths that simulate early suboptimal decisions followed by replanning. The detailed generation procedure is summarized in Algorithm 3.

C.3 Incremental Representation

To decouple transition logic from absolute node indices, we transform trajectories into an incremental format. Each path $(u_0, u_1, \ldots, u_T)$ is mapped to a sequence $[S, G, \text{inc}_1, \ldots, \text{inc}_T]$, where $\text{inc}_t = u_t - u_{t-1}$.

The vocabulary $\mathcal{V}$ is partitioned into separate segments for nodes and relative increments to prevent ID collisions. This partitioning ensures the model learns to predict the next action as a mathematical offset from the current state, rather than simply memorizing global node co-occurrences or spatial relationships. All sequences are padded to a fixed block size for efficient batch training.

C.4 Training Configurations

We train a 6-layer, 6-head, 120-dimensional Transformer for 20,000 iteration to ensure convergence, providing sufficient model capacity to master the task. Optimization is performed using the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.95$) with a global batch size of 1,024 and weight decay of 0.1. The learning rate peaks at 5×10^{-4} and follows a cosine decay schedule to 5×10^{-5} after 1,000 warmup iterations. To maintain stability, we employ `bfloat16` mixed-precision training and gradient clipping at a threshold of 1.0.

D Computational Efficiency of LSE-MTP

The procedural realization of the LSE-MTP training objective is summarized in Algorithm 4. It shares the same structure as standard MTP, follows the standard Transformer forward pass, and introduces transition layers only during training to provide supervision for future steps.

To evaluate the practical scalability of our method, we measured the training throughput of LSE-MTP, NTP, and standard MTP on an NVIDIA GeForce RTX 3090 Ti. The results show that switching from NTP to multi-token prediction ($K = 4$) leads to a 4.5% increase in parameters and a 17% drop in tokens per second, but adding the LSE constraints on standard MTP incurs almost no additional computational cost.

The training throughput of LSE-MTP remains comparable to that of standard MTP (around 425k tokens/s). This is because the latent consistency and semantic anchoring losses operate directly on the hidden representations through lightweight mean squared error (MSE) computations, which are far less expensive than the linear projections and vocabulary-wide classification heads required by standard MTP. Moreover, all auxiliary heads are discarded after training, so LSE-MTP introduces no additional latency during inference. These find-

Algorithm 3 Diverse Path Generation

Require: Graph G, pairs $\mathcal{P}_{train}$, K, p_{detour}, p_{rec}
Ensure: Training Dataset $\mathcal{D}$

$\mathcal{D} \leftarrow \emptyset$
for each pair $(s, g) \in \mathcal{P}_{train}$ **do**
  $p^* \leftarrow$ shortest path from s to g
  $\mathcal{P}_K \leftarrow$ top K shortest paths from s to g
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{[s, g, p] \mid p \in \mathcal{P}_K\}$
  if rand() $<$ p_{detour} **and** length(p^*) ≥ 4 **then**
    $v_{obs} \leftarrow$ random node in $p^* \setminus \{s, g\}$
    $G' \leftarrow G \setminus \{v_{obs}\}$
    $p_{det} \leftarrow$ shortest path from s to g in G'
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{[s, g, p_{det}]\}$ ▷ if path exists
  end if
  if rand() $< p_{rec}$ **then**
    $v_{next} \leftarrow$ second node in p^*
    $v_{wrong} \leftarrow$ random neighbor of s s.t. $v_{wrong} \neq v_{next}$
    $p_{rec} \leftarrow$ shortest path from v_{wrong} to g
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{[s, g, (s) \oplus p_{rec}]\}$
  end if
end forreturn $\mathcal{D}$

ings indicate that, once the infrastructure for multi-step supervision is in place, LSE-MTP provides an almost "cost-free" mechanism to bridge the gap between discrete token prediction and continuous world modeling.

E Further Analysis of LSE-MTP

We generated new ER and USG graphs and reduced the training set to 50% of reachable node pairs. Under the same training path generation setup, we evaluated all models on previously unseen path planning tasks.

Overall Navigation Accuracy. Table 5 shows that with moderate LSE-MTP hyperparameters, performance consistently improves over the corresponding MTP models across planning horizons K. Setting λ_s to zero noticeably degrades performance, underscoring the role of semantic alignment. As K increases, navigation accuracy declines due to reliance on next-step predictions, and adding future prediction losses can weaken this next-step capability by forcing trade-offs between objectives.

Preserving Latent Space Discriminability. Table 6 shows that semantic anchoring (λ_s) preserves latent space discriminability. Without it, the latent space collapses and topologically distinct nodes become indistinguishable. Incorporating $\mathcal{L}_{semantic}$ keeps path alignment within a meaningful manifold and prevents representational collapse.

Suppressing Structural Hallucinations. Table 7 shows that LSE reduces illegal shortcut paths (ISP). Standard MTP models with only distant supervision often skip intermediate steps in latent space. Anchoring predictions to intermediate hidden states enforces step-by-step transitions, reducing ISP and ensuring valid path connectivity.

Algorithm 4 Latent Semantic Enhancement MTP

Require: Input sequence $U = \{u_1, \ldots, u_T\}$; Prediction horizon K; Weights λ_l, λ_s.
Ensure: Total training loss $\mathcal{L}_{total}$.

$\mathbf{H} \leftarrow \text{BackboneTransformer}(U)$ ▷ $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$
$\mathcal{L}_{NTP} \leftarrow \text{CrossEntropy}(\text{Head}_{shared}(\mathbf{H}), U_{target})$
$\mathcal{L}_{MTP} \leftarrow 0$
for $k = 2$ **to** K **do**
  $\hat{\mathbf{h}}_k \leftarrow \text{LinearProj}_{k-1}(\mathbf{H})$ ▷ Project current state to future latent
  $\mathcal{L}_{CE}^{(k)} \leftarrow \text{CrossEntropy}(\text{Head}_{shared}(\hat{\mathbf{h}}_k), U_{target+k-1})$
  // Latent Consistency: Align predicted latent with future ground-truth latent
  $\mathcal{L}_{latent}^{(k)} \leftarrow \text{MSE}(\hat{\mathbf{h}}_k[: T-k+1], \mathbf{H}[k:])$
  // Semantic Anchoring: Align predicted latent with target embeddings
  $\mathbf{E}_{target} \leftarrow \text{EmbeddingLayer}(U_{target+k-1}).\text{detach}()$
  $\mathcal{L}_{semantic}^{(k)} \leftarrow \text{MSE}(\hat{\mathbf{h}}_k, \mathbf{E}_{target})$
  $\mathcal{L}_{MTP} \leftarrow \mathcal{L}_{MTP} + \mathcal{L}_{CE}^{(k)} + \lambda_l \mathcal{L}_{latent}^{(k)} + \lambda_s \mathcal{L}_{semantic}^{(k)}$
end for
$\mathcal{L}_{total} \leftarrow \mathcal{L}_{NTP} + \mathcal{L}_{MTP}$
return $\mathcal{L}_{total}$

Table 5: Detailed navigation performance metrics across different horizons (K) and LSE hyperparameter configurations (λ_l, λ_s). Suc, Disc, and WT represent Success rate, Disconnection rate, and Wrong Target rate, respectively.

Model	K	**Hyperparams** (λ_l, λ_s)	**ER (Erdős–Rényi Graph)**			**USG (Urban Street Graph)**		
			Suc ↑	Disc ↓	WT ↓	Suc ↑	Disc ↓	WT ↓
NTP (1TP)	1	-	91.80	5.95	2.25	96.22	2.57	1.21
MTP		-	91.99	6.44	**1.57**	97.13	1.70	1.17
LSE-MTP		(0.1, 0.1)	**92.68**	**5.43**	1.89	97.98	1.31	0.71
LSE-MTP	2	(0.3, 0.3)	91.22	6.33	2.45	**98.36**	**1.13**	**0.51**
LSE-MTP		(0.5, 0.5)	91.05	6.66	2.30	97.92	1.33	0.75
LSE-MTP		(0.3, 0)	85.32	11.34	3.35	97.33	2.00	0.67
MTP		-	89.50	8.37	2.13	96.28	2.67	1.05
LSE-MTP		(0.1, 0.1)	**90.62**	**7.41**	**1.98**	97.19	1.58	1.23
LSE-MTP	3	(0.3, 0.3)	88.90	8.29	2.81	**97.56**	**1.47**	**0.97**
LSE-MTP		(0.5, 0.5)	87.91	9.12	2.96	97.13	2.06	0.81
LSE-MTP		(0.3, 0)	88.36	8.87	2.77	97.03	2.18	0.79
MTP		-	87.72	**9.02**	3.26	95.72	3.39	0.89
LSE-MTP		(0.1, 0.1)	**87.81**	9.53	**2.66**	97.17	2.00	0.83
LSE-MTP	4	(0.3, 0.3)	86.58	10.39	3.03	**97.29**	**1.90**	**0.81**
LSE-MTP		(0.5, 0.5)	85.12	11.16	3.71	96.75	2.04	1.21
LSE-MTP		(0.3, 0)	85.27	11.44	3.28	96.34	2.75	0.91

Table 6: Detailed representation similarity metrics across different horizons (K) and LSE hyperparameter configurations (λ_l, λ_s). G and P indicate Goal and current Position, with $=$ and $\neq$ representing identical or different conditions. Baseline refers to the $G \neq, P \neq$ condition.

Model	K	**Hyperparams** (λ_l, λ_s)	**ER (Erdős–Rényi Graph)**				**USG (Urban Street Graph)**			
			$G =, P =$	$G =, P \neq$	$G \neq, P =$	Baseline	$G =, P =$	$G =, P \neq$	$G \neq, P =$	Baseline
NTP (1TP)	1	-	0.267	0.108	0.091	0.016	0.248	0.112	0.118	0.051
MTP	2	-	0.376	0.221	0.120	0.057	0.298	0.118	0.119	0.039
LSE-MTP	2	(0.1, 0.1)	0.396	0.263	0.129	0.069	0.346	0.152	0.176	0.078
LSE-MTP	2	(0.3, 0.3)	0.382	0.264	0.122	0.068	0.366	0.160	0.200	0.093
LSE-MTP	2	(0.5, 0.5)	0.380	0.270	0.133	0.086	0.369	0.176	0.216	0.107
LSE-MTP	2	(0.3, 0)	0.676	0.612	0.475	0.438	0.814	0.724	0.736	0.681
MTP	3	-	0.410	0.281	0.136	0.085	0.317	0.118	0.121	0.027
LSE-MTP	3	(0.1, 0.1)	0.456	0.355	0.149	0.094	0.372	0.169	0.182	0.082
LSE-MTP	3	(0.3, 0.3)	0.439	0.345	0.137	0.091	0.391	0.167	0.203	0.086
LSE-MTP	3	(0.5, 0.5)	0.450	0.361	0.137	0.094	0.394	0.178	0.213	0.103
LSE-MTP	3	(0.3, 0)	0.760	0.724	0.557	0.534	0.830	0.732	0.748	0.690
MTP	4	-	0.419	0.307	0.149	0.094	0.337	0.124	0.115	0.026
LSE-MTP	4	(0.1, 0.1)	0.469	0.372	0.160	0.106	0.379	0.158	0.165	0.063
LSE-MTP	4	(0.3, 0.3)	0.480	0.398	0.165	0.118	0.407	0.176	0.208	0.098
LSE-MTP	4	(0.5, 0.5)	0.480	0.405	0.162	0.127	0.429	0.196	0.226	0.112
LSE-MTP	4	(0.3, 0)	0.812	0.787	0.650	0.632	0.825	0.725	0.742	0.686

Table 7: Detailed next-step probability coupling and structural hallucinations across different horizons (K) and LSE hyperparameter configurations (λ_l, λ_s). ISP and Legal Prob denote illegal shortcut and valid action probabilities, respectively.

Model	K	**Hyperparams** (λ_l, λ_s)	**ER (Erdős–Rényi Graph)**		**USG (Urban Street Graph)**	
			ISP ↓	Legal Prob ↑	ISP ↓	Legal Prob ↑
NTP (1TP)	1	-	7.9×10^{-5}	0.989	2.2×10^{-5}	0.999
MTP	2	-	1.22×10^{-4}	0.988	5.7×10^{-5}	**0.998**
LSE-MTP	2	(0.1, 0.1)	$\mathbf{4.6 \times 10^{-5}}$	**0.990**	$\mathbf{2.5 \times 10^{-5}}$	**0.998**
LSE-MTP	2	(0.3, 0.3)	5.2×10^{-5}	**0.990**	3.9×10^{-5}	**0.998**
LSE-MTP	2	(0.5, 0.5)	6.2×10^{-5}	0.989	3.1×10^{-5}	**0.998**
LSE-MTP	2	(0.3, 0)	7.1×10^{-5}	**0.990**	2.7×10^{-5}	**0.998**
MTP	3	-	1.10×10^{-4}	0.985	7.8×10^{-5}	0.995
LSE-MTP	3	(0.1, 0.1)	9.1×10^{-5}	**0.989**	$\mathbf{4.0 \times 10^{-5}}$	**0.997**
LSE-MTP	3	(0.3, 0.3)	$\mathbf{3.6 \times 10^{-5}}$	**0.989**	4.4×10^{-5}	**0.997**
LSE-MTP	3	(0.5, 0.5)	1.06×10^{-4}	0.987	4.6×10^{-5}	**0.997**
LSE-MTP	3	(0.3, 0)	4.9×10^{-5}	**0.989**	4.7×10^{-5}	**0.997**
MTP	4	-	1.57×10^{-4}	0.981	1.27×10^{-4}	0.991
LSE-MTP	4	(0.1, 0.1)	$\mathbf{1.03 \times 10^{-4}}$	**0.986**	$\mathbf{5.2 \times 10^{-5}}$	**0.996**
LSE-MTP	4	(0.3, 0.3)	1.85×10^{-4}	0.985	5.9×10^{-5}	**0.996**
LSE-MTP	4	(0.5, 0.5)	1.14×10^{-4}	0.985	7.0×10^{-5}	**0.996**
LSE-MTP	4	(0.3, 0)	1.33×10^{-4}	0.984	8.5×10^{-5}	0.995