Token Is All You Need: Cognitive Planning through Belief–Intent Co-Evolution

Shiyao Sang[1]

***Abstract*— We challenge the long-standing assumption that exhaustive scene modeling is required for high-performance end-to-end autonomous driving (E2EAD). Inspired by cognitive science, we propose that effective planning arises not from reconstructing the world, but from the co-evolution of belief and intent within a minimal set of semantically rich tokens. Experiments on the nuPlan benchmark (720 scenarios, 11k+ samples) reveal three principles: (1) sparse intent tokens alone achieve 0.487 m ADE, demonstrating strong performance without future prediction; (2) conditioning trajectory decoding on predicted future tokens reduces ADE to 0.382 m, a 21.6% improvement, showing that performance emerges from cognitive planning; and (3) explicit reconstruction loss degrades performance, confirming that task-driven belief–intent co-evolution suffices under reliable perception inputs. Crucially, we observe the emergence of cognitive consistency: through prolonged training, the model spontaneously develops stable token dynamics that balance current perception (belief) and future goals (intent). This process, accompanied by "temporal fuzziness," enables robustness under uncertainty and continuous self-optimization. Our work establishes a new paradigm: intelligence lies not in pixel fidelity, but in the tokenized duality of belief and intent. By reframing planning as understanding rather than reaction, TIWM bridges the gap between world models and VLA systems, paving the way for foresightful agents that plan through imagination. Note: Numerical comparisons with methods reporting results on nuScenes are indicative only, as nuPlan presents a more challenging planning-focused evaluation.**

I. Introduction

End-to-end autonomous driving (E2EAD) seeks to map perception directly to control, bypassing traditional pipelines that decompose perception, prediction, and planning [1], [2], [3]. Recent works such as SSR [4] achieve strong performance through sparse BEV representations and temporal self-supervision. Yet, most approaches still assume that effective planning requires dense scene representations and future self-supervision, leading to complex architectures that simulate full scene dynamics or rely on restrictive Markov assumptions. World-model methods attempt to generate future scenes in latent space, but incur high computational cost and error accumulation in long-horizon reasoning. Likewise, vision-language-action (VLA) frameworks [5] introduce symbolic reasoning but typically assume temporal independence, limiting their ability to handle multimodal or uncertain futures. Humans, by contrast, neither reconstruct every visual detail nor act solely on the present. They maintain compact cognitive models, selectively attending to semantically relevant cues while mentally simulating possible futures. Inspired by this, we hypothesize that high-performance planning arises not from reconstructing the world but from understanding it—through a minimal set of semantically meaningful tokens encoding navigational intent. We propose the **Tokenized Intent World Model (TIWM)**, a cognitively inspired framework that redefines future reasoning in E2EAD:

- **Sparse intent representation:** environmental semantics are compressed into a small set of high-dimensional tokens capturing task-relevant information, yielding efficiency and interpretability;
- **Task-driven semantic alignment:** the model aligns intent tokens directly with planning objectives, eliminating the need for pixel-level reconstruction;
- **Temporal fuzziness as a cognitive feature:** reasoning probabilistically about future states enables robust and flexible planning under uncertainty.
- **Cognitive consistency learning:** the system achieves persistent performance gains through prolonged training by dynamically balancing belief (current state understanding) and intent (future goal imagination), forming a self-organizing internal world model—a computational manifestation of belief–intent co-evolution.

Experiments on `nuPlan` validate these principles. Even without future prediction, our sparse model achieves 0.487 m ADE, challenging the belief that dense modeling is essential. Conditioning trajectory decoding on predicted future tokens further improves ADE to 0.382 m—a 21.6% gain within the same framework—demonstrating that performance emerges from *semantic imagination* rather than pixel reconstruction. Adding explicit reconstruction loss provides no benefit, confirming that task-driven semantic alignment alone suffices under reliable perception. Together, these results reveal a paradigm shift: effective planning stems not from exhaustive reconstruction but from selective understanding. By treating *intent* as the atomic unit of world representation, TIWM bridges two previously disjoint paradigms—world models that simulate and VLA systems that interpret—showing that **belief–intent co-evolution**, rather than dense reconstruction, underlies human-like planning.

II. Related Work

Autonomous driving has evolved from modular pipelines toward end-to-end (E2E) frameworks. Early systems decomposed perception, prediction, and planning into separate stages—interpretable but prone to error propagation and

*This work was conducted independently, without institutional or external funding support.

[1]Shiyao Sang is an independent researcher, Jiangsu, China. www.fengmao@outlook.com

hand-crafted heuristics [6], [7]. Modern E2E approaches employ deep neural networks to map sensory inputs directly to control signals, enabling joint optimization across the full decision-making stack [1], [2], [3], [4].

Vision-based methods such as UniAD [1] and VAD [2] achieve strong results via multi-camera fusion and multi-task learning. However, their reliance on dense feature reconstruction and short-horizon prediction limits long-term reasoning and scalability. In contrast, our work demonstrates that high-performance planning can emerge from a minimal set of semantically rich tokens—without dense reconstruction or auxiliary supervision.

Latent world modeling frameworks such as *World Models* [8] and *Dreamer* [9] enable "mental simulation" through compact latent dynamics. Yet, they face two major limitations in real-world driving: (1) the computational burden of generating full future scenes, and (2) the absence of theoretical guarantees for modeling multimodal behaviors, which leads to compounding errors in closed-loop control [10]. Our experiments show that this reconstruction is unnecessary—aligning sparse, task-relevant semantics with future intents suffices for high-quality planning. Unlike these scene-generative approaches, we focus solely on decision-critical information, showing that *"token is all you need"* for world modeling.

Vision-language-action (VLA) frameworks unify perception, reasoning, and control [5], often conditioning on natural-language commands to enhance interpretability. However, most assume temporal independence between observations, producing reactive rather than anticipatory policies.

Sparse scene modeling offers an alternative. SSR [4] extracts 16 navigation-guided tokens from BEV features and achieves 0.75 m ADE on `nuScenes`, using a Future Feature Predictor to align predicted and real BEV features via L2 loss. While conceptually aligned with our approach, SSR still relies on dense BEV feature reconstruction for self-supervision. As its authors note, "we supervise directly using an L2 loss with the real future BEV feature," an assumption our results refute. We show that reconstruction loss not only fails to improve performance but can impede the emergence of high-level intent representations, whereas task-driven semantic alignment yields superior results.

Most E2E planners—whether perception-to-trajectory regressors or VLM/VLA hybrids [11], [12]—remain constrained by Markovian dynamics, neglecting temporal dependencies and structural similarities in high-dimensional state spaces. This often leads to multi-objective conflicts and degraded safety generalization [10]. Consequently, such systems lack genuine "mental simulation," i.e., the capacity to anticipate multiple futures before acting. Our method introduces autoregressive intent chains that enable future-oriented reasoning through task-driven semantic alignment rather than pixel-perfect reconstruction, capturing the essence of cognitive planning while avoiding over-constrained temporal alignment.

Drawing from cognitive science, prior studies have explored attention, memory, and imagination in autonomous decision-making [13], [14], [15]. However, most remain conceptual or limited to architectural analogies. We instead operationalize two core cognitive principles—*mental simulation* [16], [17] and *sparse coding* [18]—within a unified generative–planning framework. This integration yields a computationally efficient yet cognitively grounded planner that achieves state-of-the-art performance.

In summary, while existing E2E and world-model approaches either reconstruct dense futures or assume Markovian independence, our **Tokenized Intent World Model (TIWM)** demonstrates that effective planning can emerge from sparse, semantically aligned intent tokens. This reframes the goal of autonomous reasoning—from *reconstructing the world* to *understanding it*—enabling agents that plan through imagination rather than reaction.

III. Method

We propose the **Tokenized Intent World Model (TIWM)**, a cognitively grounded planning framework that redefines end-to-end autonomous driving (E2EAD) as a process of *belief–intent co-evolution* (Fig. 1). Unlike conventional approaches that rely on dense feature maps or pixel-level future reconstruction, TIWM achieves high-performance planning through a compact set of semantically rich tokens, requiring neither full-scene generation nor auxiliary self-supervision.

A. Sparse Intent Representation

TIWM operates on perception-informed Bird's Eye View (BEV) representations, structured as a grid with two channels for dynamic objects and four for semantic map layers (`LANE`, `INTERSECTION`, `STOP_LINE`, `CROSSWALK`). Following standard practice, we assume access to reliable BEV inputs from a perception stack, thereby isolating the study of decision-level behavior.

The core component is the *Sparse Token Learner*, which compresses a dense BEV tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ into a compact set of $N{=}16$ tokens $\mathbf{T} \in \mathbb{R}^{N\times D}$:

$$\mathbf{T} = \sum_{hw} \text{Softmax}(\phi(\mathbf{X})) \odot \mathbf{X}, \tag{1}$$

where the attention weights are computed via a lightweight CNN:

$$\phi(\mathbf{X}) = \text{Conv}_2(\text{ReLU}(\text{Conv}_1(\mathbf{X}))). \tag{2}$$

This process condenses a $256{\times}256$ BEV grid into 16 high-level semantic tokens, each representing a critical aspect of the driving scene. This design operationalizes the cognitive principle of *sparse coding* [18], where minimal active units capture the most decision-relevant information.

B. Intent Chain Evolution via Belief–Intent Co-Evolution

TIWM models planning as the dynamic interplay between *belief* (understanding of the present) and *intent* (imagination of the future). Instead of predicting full future states, the model autoregressively evolves an *intent chain* from a sequence of historical tokens.

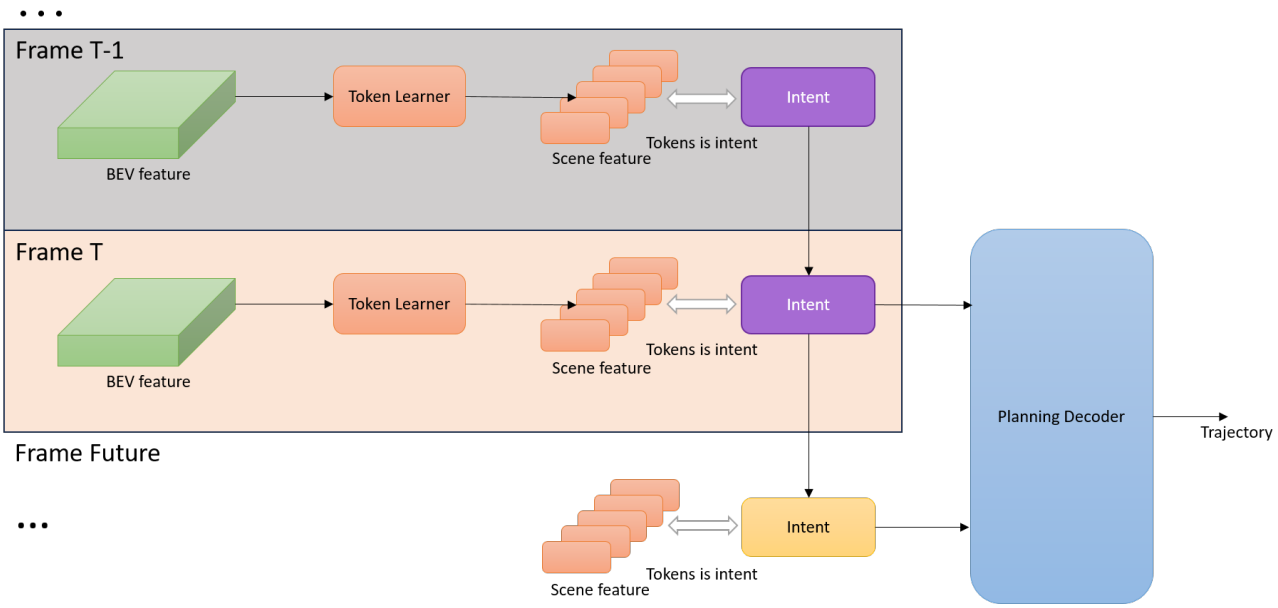

Fig. 1. **Tokenized Intent World Model: from perception to cognitive world.** At each timestep, sparse *tokens* are extracted from BEV perception features, representing the distilled semantics of the scene. The model then autoregressively predicts future intent tokens—compact, task-relevant abstractions of the agent's imagined goals. The planning decoder jointly reasons over current sparse tokens and predicted future intents to generate executable trajectories. Each intent token serves three roles: (1) **Representation** — compressing the core cognitive state of the environment; (2) **Intent** — unfolding plausible futures through autoregressive prediction; (3) **Decision** — directly guiding the trajectory decoder toward goal-directed action. In this framework, the *world* becomes an actionable cognitive entity—interpreted, imagined, and utilized by the decision system—rather than a dense intermediate reconstruction.

Given four historical token sets $\mathbf{T}_{t-3:t}$, a learnable future query $\mathbf{Q} \in \mathbb{R}^{N\times D}$, and a causal attention mask $\mathcal{M}$, we compute the future intent $\mathbf{I}_{t+1}$ as:

$$[\mathbf{T}_{t-3:t}, \mathbf{I}_{t+1}] = \text{TransformerEncoder}(\text{Concat}(\mathbf{T}_{t-3:t}, \mathbf{Q}); \mathcal{M}), \tag{3}$$

The TransformerEncoder consists of two layers with eight attention heads and an embedding dimension of $D = 256$. The causal mask $\mathcal{M}$ ensures that future intent tokens can only attend to past and present information, preventing information leakage.

This architecture enables **belief–intent co-evolution**: the current tokens $\mathbf{T}_t$ form the belief state, while the predicted $\mathbf{I}_{t+1}$ constitutes the intent state. They are not independent; rather, they are jointly shaped by a single task-driven objective, leading to a coherent internal world model.

C. Task-Driven Semantic Alignment and Cognitive Consistency

TIWM's training objective is intentionally minimalist, relying solely on the final planning outcome:

$$\mathcal{L}_{\text{traj}} = \lambda_{\text{traj}} \cdot \text{SmoothL1}(\text{TrajectoryPredictor}(\mathbf{T}_t, \mathbf{I}_{t+1}), \mathbf{Y}), \tag{4}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{traj}}, \tag{5}$$

where $\mathbf{Y}_{t+1:t+H}$ is the ground-truth 3-second trajectory ($H = 6$).

For ablation, we also evaluate an auxiliary reconstruction loss that aligns the predicted intent with a ground-truth token:

$$\mathcal{L}_{\text{intent}} = \lambda_{\text{intent}} \| \text{F.normalize}(\mathbf{I}_{t+1}) - \text{F.normalize}(\mathbf{T}^{\text{gt}}_{t+1}) \|_2, \tag{6}$$

with $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{intent}}$. However, this explicit supervision consistently degrades performance, confirming that trajectory-driven learning alone is sufficient—and superior—for inducing meaningful intent representations.

Through prolonged training, this task-driven process leads to the emergence of **cognitive consistency**: the belief–intent pair self-organizes into a stable, robust equilibrium. This is not a static state but a dynamic balance, akin to a "quantum state transition" between representing "what is" and "what could be," enabling persistent self-optimization over thousands of epochs.

D. Cognitive Planning Mechanism

The final trajectory is generated by a Multi-Modal Trajectory Predictor:

$$\mathbf{Y}_{t+1:t+H} = \text{TrajectoryPredictor}(\mathbf{T}_t, \mathbf{I}_{t+1}). \tag{7}$$

The predictor first computes mode weights $\mathbf{w} = \text{softmax}(\mathbf{W}_m \mathbf{I}_{t+1})$, then fuses current context and intent via cross-attention, and finally decodes the fused representation into 2D trajectory points. This process embodies *planning through imagination*, where the agent reasons over possible futures before committing to an action.

IV. Experiments

A. Experimental Setup

Dataset. We evaluate the proposed Tokenized Intent World Model (TIWM) on the `nuPlan` benchmark for urban autonomous driving. Following standard practice, we use 720 scenarios (≈11k samples) with a 9:1 train–validation split, covering diverse settings such as intersections, merges, and multi-agent interactions.

Input Representation. We employ NuPlan's `RasterFeatureBuilder` to construct Bird's-Eye-View (BEV) inputs from structured perception and HD map data. Four semantic layers (`LANE`, `INTERSECTION`, `STOP_LINE`, `CROSSWALK`) and two vehicle-occupancy channels are included, providing clean geometry aligned with real-world deployment where perception modules are reliable.

Evaluation Metrics. Performance is measured by Average Displacement Error (ADE) and Final Displacement Error (FDE) over a 3 s horizon at 0.5 s intervals, following the standard `MAX` protocol used in UniAD and SSR. ADE reflects mean trajectory deviation, while FDE captures the terminal-point error.

Implementation Details. TIWM adopts a ResNet-18 encoder (d=256) and processes four historical frames (2 s) to predict six future points (3 s). Training uses AdamW (β_1=0.9, β_2=0.999), batch size 32, learning rate 1×10^{-4} with cosine decay, and weight decay 1×10^{-4}. All models are trained on a single NVIDIA RTX 4090 (24 GB).

B. Main Results

Table I summarizes our ablation results on `nuPlan`, along with representative state-of-the-art (SOTA) methods on `nuScenes`. TIWM achieves strong performance through *task-driven semantic alignment*, without dense reconstruction or auxiliary supervision.

a) Key Findings.: **(1) Minimal complexity yields strong performance.** Even without future prediction, TIWM achieves 0.487 m ADE, comparable to or surpassing prior benchmarks on `nuScenes`. This demonstrates that semantically reliable inputs enable high-quality planning without dense reconstruction.

(2) Future-conditioned decoding enhances reasoning. Conditioning trajectory generation on predicted future tokens—rather than reconstructed BEV features—reduces ADE to 0.382 m, a 21.6% gain over the baseline. This validates that planning via semantic imagination, rather than pixel-level reconstruction, drives performance.

(3) Explicit reconstruction loss hinders learning. Incorporating future token decoding alongside an intent-alignment term slightly degrades performance (ADE: 0.492 m vs. 0.382 m) and introduces instability in validation performance. Furthermore, this configuration yields a higher training loss compared to the other three variants, suggesting a gradient conflict arising from the competing objectives of trajectory reconstruction and intent alignment.

b) Interpretation.: The key distinction between TIWM and SSR lies in their self-supervision philosophy. SSR [4] enforces pixel-level alignment by reconstructing future BEV features under auxiliary trajectory guidance, effectively coupling perception, prediction, and planning through dense supervision. In contrast, TIWM adopts a task-driven, token-level semantic alignment strategy—reconstructing only the compact intent representation that is directly optimized for the planning objective. This design decouples low-level feature reconstruction from high-level decision semantics, resulting in a more efficient and cognitively interpretable learning process.

C. Ablation Study

To assess each component's role, we evaluate four controlled variants:

- **Current tokens without intent loss:** decode from current tokens only, without future prediction or reconstruction.
- **Current tokens with intent loss:** decode from current tokens with auxiliary reconstruction supervision. $\mathcal{L}_{\text{intent}} = 100$;
- **Future tokens without intent loss:** decode from current and predicted future tokens without reconstruction.
- **Future tokens with intent loss:** decode from current and predicted future tokens with explicit L2 reconstruction.$\mathcal{L}_{\text{intent}} = 100$;

Fig. 2 reports validation ADE and training loss across configurations. All models improve steadily up to epoch 100, demonstrating strong potential for further convergence. However, prolonged training incurs significant computational costs. This trend suggests that longer training enhances temporal reasoning, albeit at the expense of efficiency.

The contrast between *Future tokens, no intent loss* and *Future tokens, with intent loss* is particularly revealing. Both condition the decoder on predicted intents, yet only the former allows trajectory supervision to structure the latent token space. The reconstruction loss biases tokens toward low-level similarity, suppressing the emergence of high-level semantic alignment essential for effective planning.

D. Convergence Study

As shown in Table II, the model's performance improves with increasing training epochs, albeit with diminishing returns. The consistent downward trend in both validation ADE and training loss (Fig. 3) suggests that the model possesses a remarkable capacity for prolonged learning, continuously refining its internal representations even after 1000 epochs.

This persistent convergence is not merely a numerical phenomenon; it reflects a deeper principle of *cognitive consistency*. We propose that the sparse tokens within the intent chain function as high-level cognitive units, embodying a dual nature:

- **Belief**: They serve as a compressed representation of the current environment, encoding task-relevant semantics.
- **Intent**: They represent the agent's imagined future goals, guiding the planning process.

Table I: *Ablation results on* `nuPlan`*, with SOTA methods on* `nuScenes` *shown for reference.* Our framework achieves high accuracy via sparse intent alignment, without explicit reconstruction. Numerical comparisons across datasets are indicative only.

Method	Future Intent	Self-Supervision Type	ADE (m)	FDE (m)	Params (M)
Methods evaluated on nuScenes (reference only)					
UniAD [1]	×	Multi-task Auxiliary	1.03	1.65	∼80
VAD-Base [2]	×	Multi-task Auxiliary	1.22	1.98	∼55
SSR [4]	×	Dense Reconstruction	0.75	1.36	∼32
Our ablation study on nuPlan (main results)					
Current token without intent loss	×	None	0.487	0.855	14.2
Current token with intent loss	×	Sparse Intent Reconstruction	0.459	0.810	14.2
Future token without intent loss	✓	Task-driven Semantic Alignment	**0.382**	**0.699**	14.2
Future token with intent loss	✓	Sparse Intent Reconstruction	0.492	0.944	14.2

Note: Parameter counts for SSR, UniAD, and VAD are estimated since exact values are not reported. SSR uses dense BEV reconstruction ("we supervise directly using an L2 loss with the real future BEV feature" [4]), whereas TIWM aligns only 16 sparse intent tokens—conceptually distinct from feature-level reconstruction. Within `nuPlan`, the *Future token without intent loss* setting attains 0.382 m ADE—16.8% better than its reconstruction-based variant with current token.

Dataset Context: `nuScenes` and `nuPlan` share a common data origin but serve different purposes. `nuPlan` is specifically designed as a planning benchmark, featuring complex urban scenarios curated to stress-test long-horizon reasoning and interactive decision-making, making it more challenging than `nuScenes` for evaluating pure planning performance. In contrast, many two-stage E2E methods on `nuPlan` report closed-loop metrics that often rely on rule-based post-processing to ensure safety, obscuring their open-loop trajectory prediction capabilities. Our results on `nuPlan` are reported in open-loop ADE/FDE, providing a clean evaluation of the core planning module without such post-processing.

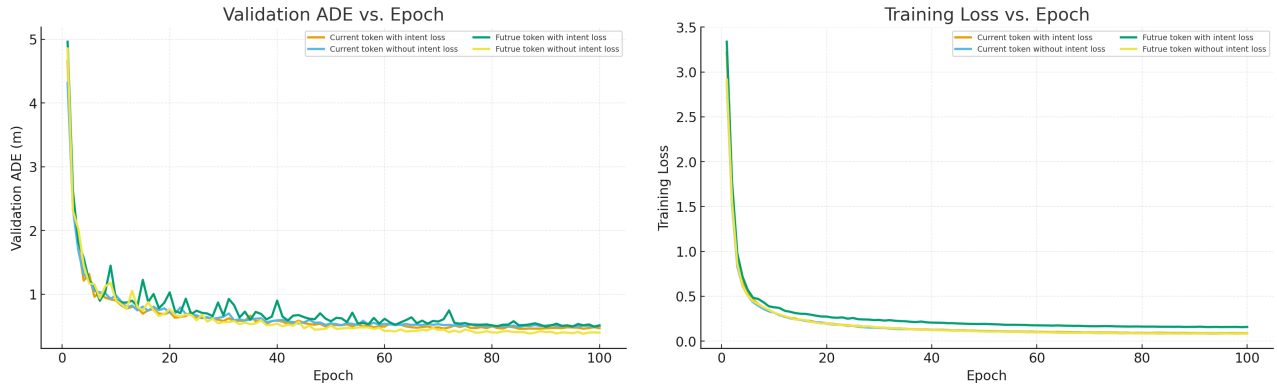

Fig. 2. Validation ADE (left) and training loss (right) versus epoch across four configurations. All runs converge stably, with late-epoch best ADEs between 0.382 and 0.492 m. Early-epoch behavior is configuration-dependent: with current tokens, intent conditioning yields modest early gains (lower mean ADE within the first 100 epochs), whereas with future tokens, the non-intent variant reaches lower ADE faster. The overall best minimum is obtained by the future-token model without intent.

We hypothesize that during training, these token units undergo a dynamic equilibrium between their "belief" and "intent" states—a process we metaphorically describe as a "quantum state transition." This continuous balancing act represents a more fundamental form of learning than conventional supervised or reinforcement learning, as it involves the self-organization of an internal world model.

This interpretation also offers a compelling explanation for the biological phenomena of neural replay [19] and memory consolidation [20]: just as the brain replays past experiences to strengthen memories and refine decision-making policies, our model's long-term training can be seen as a form of "computational replay," where the system repeatedly refines its internal belief-intent space to achieve greater robustness and generalization.

The convergence curve thus reveals not just a decline in error, but the emergence of an implicit, natural regularization within the system—a self-organizing process that embodies the very essence of adaptive, cognition-inspired intelligence. This is a emergence of cognitive intelligence.

V. Discussion

Our experiments reveal four foundational principles of **cognitive planning** that collectively challenge the dominant paradigms in end-to-end autonomous driving.

(1) Task-driven semantic alignment surpasses pixel-level reconstruction. The consistent superiority of the *Future token without intent loss* configuration (0.382 m ADE) over its reconstruction-based counterpart (0.492 m ADE) demonstrates that explicit alignment with future BEV features is not only unnecessary but detrimental. This aligns with human cognition, where attention prioritizes task-relevant cues rather than exhaustively reconstructing the visual scene. As SSR [4] notes, reconstruction-heavy paradigms often couple perception, prediction, and planning through dense supervision, leading to unstable convergence and entangled objectives [10]. By focusing learning solely on the final planning outcome, TIWM decouples low-level feature generation from high-level decision semantics, yielding a more stable, efficient, and interpretable system.

(2) "Temporal fuzziness" is a cognitive feature, not a flaw. The model learns *what* to focus on rather than

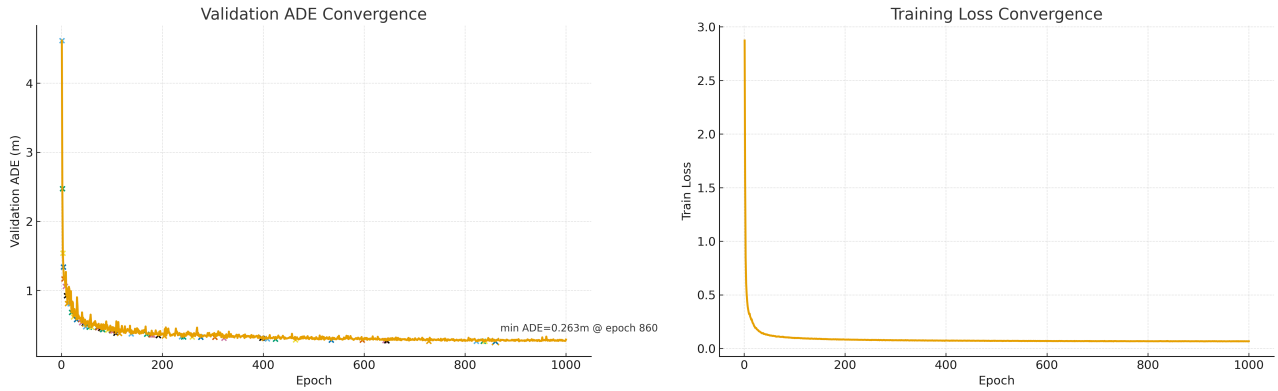

Fig. 3. Training dynamics for the "Future token without intent loss" configuration. **Left**: Validation ADE (m) converges to a minimum of 0.263 m at epoch 860. **Right**: Training loss decreases steadily with epoch, indicating continuous optimization. This sustained convergence demonstrates the model's capacity for prolonged learning and its potential for further performance gains.

Table II: Performance Evolution of TIWM under Prolonged Training.

Training Epochs	Future Intent	Self-Supervision Type	ADE (m)	FDE (m)	Improvement vs 50ep (%)
50	✓	Task-driven Semantic Alignment	0.431 ± 0.641	0.798 ± 1.199	-
100	✓	Task-driven Semantic Alignment	0.382 ± 0.584	0.699 ± 1.098	11.4%
300	✓	Task-driven Semantic Alignment	0.299 ± 0.541	0.535 ± 1.009	30.6%
500	✓	Task-driven Semantic Alignment	0.281 ± 0.546	0.506 ± 1.017	34.8%
1000	✓	Task-driven Semantic Alignment	0.263 ± 0.558	0.471 ± 1.013	39.0%
3000	✓	Task-driven Semantic Alignment	**0.262** ± 0.567	**0.463** ± 1.022	**39.2%**

Note: This table quantifies the model's exceptional capacity for continuous learning over extended training. The consistent improvement from 50 to 3000 epochs on the nuPlan validation set validates our core hypothesis: that task-driven semantic alignment enables deep, persistent refinement of internal representations, leading to superior generalization. The diminishing returns after 1000 epochs suggest a natural plateau, yet the system remains far from convergence, highlighting its robustness and adaptability.

when, maintaining probabilistic attention over key semantic elements across time. This flexibility enables robust reasoning under uncertainty and avoids overfitting to strict temporal alignment—an inherent limitation of frame-based world models. Temporal fuzziness is not noise; it is an adaptive mechanism that mirrors the brain's ability to reason about events in a temporally coarse manner, prioritizing semantic relevance over precise timing.

(3) Planning through imagination outperforms reactive policies. The 21.6% performance gain from future-conditioned decoding validates that intelligence arises not from reacting to the present, but from anticipating multiple futures before acting. This directly addresses the Markovian limitation of VLA systems [12], demonstrating that autoregressive intent chains provide a computational form of *mental simulation* [16], [17] that supports foresightful planning.

(4) Cognitive consistency learning enables persistent self-optimization. The model's continuous performance gains over 3000 epochs—converging to 0.262 m ADE—reveal an emergent self-organizing process. We propose that sparse tokens function as high-level cognitive units, dynamically balancing **belief** (current perception) and **intent** (future imagination). This equilibrium, which we metaphorically describe as a "quantum state transition," mirrors biological phenomena like neural replay [19] and memory consolidation [20]. Long-term training thus acts as "computational replay," where the system repeatedly refines its internal belief–intent space to achieve greater robustness—a hallmark of adaptive, cognition-inspired intelligence.

On Cross-Dataset Comparisons. While our results are reported on `nuPlan` and prior SOTA methods on `nuScenes`, the comparison is meaningful. `nuPlan` is explicitly designed as a *planning-focused* benchmark, featuring complex, interactive scenarios curated to stress-test long-horizon reasoning, making it more challenging than `nuScenes` for evaluating pure planning performance. In contrast, many two-stage E2E methods on `nuPlan` report closed-loop metrics that often rely on rule-based post-processing to ensure safety, obscuring their open-loop trajectory prediction capabilities. Our open-loop ADE/FDE on `nuPlan` provides a cleaner evaluation of the core planning module without such post-processing.

Together, these insights mark a paradigm shift: from **reconstructing the world** to **understanding it**. TIWM achieves stronger results not by simulating scenes more faithfully, but by aligning semantics more intelligently. This bridges two long-divided paradigms in embodied AI: world models that emphasize exhaustive synthesis, and VLA systems that are limited to instantaneous reasoning. We show that **belief–intent co-evolution**, not pixel fidelity, underpins effective planning.

Broader Implications. The "token is all you need" principle transcends autonomous driving. In robotic manipulation, sparse intent tokens could represent dynamic affordances rather than full object geometry. In conversational AI, they could encode evolving discourse states, enabling true foresightful interaction. More broadly, intelligence emerges not from reconstructing environmental details, but from the self-

organized balance between *what is* (belief) and *what could be* (intent).

Thus, TIWM is not merely a planner—it is a computational model of cognition, where tokens act as atomic units of thought. This work marks a transition from the era of computational intelligence to the era of **cognitive intelligence**.

VI. Conclusion

We challenge the prevailing dogma that high-fidelity future scene reconstruction is indispensable for effective end-to-end autonomous planning. Through extensive experiments on the `nuPlan` benchmark (720 scenarios, $\sim$11k samples), we demonstrate that a minimalist, cognitively inspired framework—operating on only 16 sparse intent tokens—can achieve state-of-the-art performance. Our Tokenized Intent World Model (TIWM) attains an ADE of **0.382 m** without any auxiliary reconstruction supervision, validating that task-driven semantic alignment is not only sufficient but superior to dense modeling paradigms.

This work establishes a new paradigm in autonomous intelligence: **planning through understanding**. We show that:

1) **High-quality perception enables powerful yet simple planners.** Reliable BEV inputs allow the system to focus on decision-critical semantics, rendering complex reconstruction losses redundant.
2) **Future-aware decoding enables foresightful behavior.** Conditioning trajectory generation on autoregressively predicted intent tokens yields a **21.6%** performance gain, illustrating that intelligence arises from *imagining possible futures*, not merely reacting to the present.
3) **Explicit reconstruction loss is not only unnecessary—it is detrimental.** It introduces gradient conflicts and hinders the emergence of high-level semantic representations, whereas a clean, trajectory-only objective naturally fosters robust and generalizable planning.

Crucially, our model exhibits an emergent cognitive property: **temporal fuzziness**—the ability to attend flexibly to *what matters* semantically, rather than aligning rigidly to *when* it occurs. This, combined with the discovery of **cognitive consistency** through prolonged training, reveals that intelligence stems not from pixel-level fidelity, but from the **coherence of internal cognitive dynamics**. The sparse token, in its dual role as belief and intent, acts as a fundamental unit of thought—a dynamic entity undergoing continuous "quantum state transition" between perceiving the world and imagining its futures.

Beyond autonomous driving, TIWM suggests a universal principle for embodied intelligence: **true autonomy arises not from reconstructing the world, but from understanding its semantic essence and simulating its possibilities.** By unifying representation, intention, and decision into a single, tokenized loop, our framework bridges the gap between data-driven perception and cognitively inspired reasoning.

Future work will explore closed-loop evaluation, scalability to multi-intent scenarios, and the generalization of belief–intent co-evolution to robotic manipulation and interactive AI. We view TIWM not merely as a planning module, but as a foundational step toward a new generation of agents that do not just *react*, but truly *plan through understanding*—ushering in the era of cognitive intelligence.

References

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-Oriented Autonomous Driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[2] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "VAD: Vectorized Scene Representation for Efficient Autonomous Driving," Aug. 2023.

[3] W. Zheng, R. Song, X. Guo, C. Zhang, and L. Chen, "GenAD: Generative End-to-End Autonomous Driving," Apr. 2024.

[4] P. Li and D. Cui, "Navigation-Guided Sparse Scene Representation for End-to-End Autonomous Driving," Mar. 2025.

[5] X. Zhou, X. Han, F. Yang, Y. Ma, and A. C. Knoll, "OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model," Mar. 2025.

[6] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Berlin, Heidelberg: Springer-Verlag, Oct. 2022, pp. 1–18.

[7] S. Shi, L. Jiang, D. Dai, and B. Schiele, "MTR++: Multi-Agent Motion Prediction With Symmetric Scene Modeling and Guided Intention Querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3955–3971, May 2024.

[8] D. Ha and J. Schmidhuber, "Recurrent World Models Facilitate Policy Evolution," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[9] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to Control: Learning Behaviors by Latent Imagination," Mar. 2020.

[10] Y. Zheng, J. Li, D. Yu, Y. Yang, S. E. Li, X. Zhan, and J. Liu, "Safe Offline Reinforcement Learning with Feasibility-Guided Diffusion Model," Jan. 2024.

[11] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models," Jun. 2024.

[12] M. Bansal, A. Krizhevsky, and A. Ogale, "ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst," Dec. 2018.

[13] M. Da Lio, A. Mazzalai, D. Windridge, S. Thill, H. Svensson, M. Yüksel, K. Gurney, A. Saroldi, L. Andreone, S. R. Anderson, and H.-J. Heich, "Exploiting dream-like simulation mechanisms to develop safer agents for automated driving: The "Dreams4Cars" EU research and innovation action," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Oct. 2017, pp. 1–6.

[14] S. Mahmoud and H. Svensson, "Self-driving cars learn by imagination," in *Swecog 2018, the 14th Swecog Conference, Linköping, Sweden, October 11-12, 2018*. University of Skövde, 2018, pp. 12–15.

[15] A. Mazzalai, "A Cognitively Inspired Framework to Support the Driving Task of Vehicles of the Future," 2018.

[16] J. Decety and J. Grèzes, "The power of simulation: Imagining one's own and other's behavior," *Brain Research*, vol. 1079, no. 1, pp. 4–14, Mar. 2006.

[17] M. G. Mattar and M. Lengyel, "Planning in the brain," *Neuron*, vol. 110, no. 6, pp. 914–934, Mar. 2022.

[18] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, Jun. 1996.

[19] Q. Huang, Z. Xiao, Q. Yu, Y. Luo, J. Xu, Y. Qu, R. Dolan, T. Behrens, and Y. Liu, "Replay-triggered brain-wide activation in humans," *Nature Communications*, vol. 15, no. 1, p. 7185, Aug. 2024.

[20] L. R. Squire, L. Genzel, J. T. Wixted, and R. G. Morris, "Memory Consolidation," *Cold Spring Harbor Perspectives in Biology*, vol. 7, no. 8, p. a021766, Jan. 2015.