Planning in 16 Tokens: A Compact Discrete Tokenizer for Latent World Model

Dongwon Kim[1] Jinsung Lee[1] Gawon Seo[1] Minsu Cho[1,2] Suha Kwak[1]
[1] POSTECH [2] RLWRLD
{kdwon, jinsunglee, gawon, mscho, suha.kwak}@postech.ac.kr

Abstract

World models provide a powerful framework for simulating environment dynamics conditioned on actions or instructions, enabling downstream tasks such as action planning or policy learning. Recent approaches leverage world models as learned simulators, but its application to decision-time planning remains computationally prohibitive for real-time control. A key bottleneck lies in latent representations: conventional tokenizers encode each observation into hundreds of tokens, making planning both slow and resource-intensive. To address this, we propose CompACT, a discrete tokenizer that compresses each observation into just 16 tokens, drastically reducing computational cost while preserving essential information for planning. An action-conditioned world model that occupies CompACT tokenizer achieves competitive planning performance with orders-of-magnitude faster planning, offering a practical step toward real-world deployment of world models.

1. Introduction

Humans navigate the world not through pixel-perfect recall of their surroundings, but rather through compact mental representations that capture only the information necessary for decision-making [19, 23]. This internal model—an imprecise but efficient abstraction of reality—reduces the complexity of sensory input into a representation optimized for action and planning. In the context of artificial intelligence and reinforcement learning, this concept manifests as the *world model* [23], a neural network that captures environment dynamics to enable planning [3, 27, 44, 69] and policy learning [1, 24–26, 62].

World models have emerged as a promising solution to the sample inefficiency of reinforcement learning (RL). Traditional model-free RL methods require millions of interactions with the environment to learn effective policies, making them impractical for real-world applications where data collection is expensive or risky. By learning to predict future states, world models enable agents to simulate experiences internally, reducing the need for real environment interactions. Furthermore, these models themselves can be used for planning without additional learning of policy [3, 69] through model-predictive control (MPC) [12, 61].

Recent advances in world modeling have been driven by the rapid progress in generative models, particularly in image and video generation [7, 16, 50]. These models can generate photorealistic images or video conditioned on language instructions [15, 62] or actions [1, 3, 62, 69, 70], suggesting implicit understanding of world's underlying dynamics.

However, there exists a critical discrepancy between generative approaches and the requirements for effective planning: *high-fidelity generation does not translate to better decision-making* [57]. To achieve photorealistic quality, these models must capture extensive perceptual detail—textures, lighting, shadows—that is largely irrelevant for action selection. This necessitates encoding single images into hundreds of latent tokens, which sharply increases computational cost. Since most world models in the literature adopt attention-based architectures [49], this burden grows quadratically, making planning especially expensive. As a result, current world models remain impractical for real-world control: for example, state-of-the-art navigation world models [3] require up to 2 minutes of computation per episode for planning, making them unsuitable for applications demanding real-time responsiveness.

We propose CompACT, a compact tokenizer that encodes each image into just 16 tokens—approximately 200 bits per image. This represents an extreme compression ratio compared to existing approaches. For instance, the SD-VAE tokenizer [50] used in NWM [3] requires 196 tokens to represent the same image. Beyond the reduction in token count, our tokenizer further distinguishes itself by employing a discrete latent space, enabling much faster future-state prediction: each token is unmasked only once [7], rather than being processed through hundreds of iterative denoising steps typically required in diffusion models utilizing continuous latent space [30].

While such extreme compression inevitably sacrifices fine-grained visual details, our tokenizer preserves low-

frequency features—high-level semantics and spatial relationships—that are crucial for planning and decision-making. The key technical contribution enabling this extreme compression is our generative decoding approach: rather than attempting direct pixel reconstruction from 16 tokens, our decoder learns to unmask the latent representation of a pretrained tokenizer, using the compact tokens as conditioning. This formulation transforms an intractable decompression problem into a tractable conditional generation task. By training world models in this compact latent space, we achieve order-of-magnitude reductions in rollout latency.

To validate the effectiveness of the proposed approach, we train NWM [3], an action-conditioned world model for navigation, on the latent space of CompACT. Such action-conditioned world models have a unique strength in that they can serve as general-purpose planners via MPC, but the prohibitive computational burden required for rollouts has remained as a bottleneck. On navigation planning in RECON [51], the NWM trained with our CompACT achieves comparable accuracy to the one using 196 continuous tokens while delivering approximately 20× speedup in planning latency. Furthermore, our 16-token model outperforms the FlexTok [2] with 64 tokens, validating that carefully designed extreme compression can yield both computational efficiency and superior planning performance.

2. Method

2.1. Latent generative model as world model

In this section, we first describe how a world model can be formulated as latent generative models. We consider the standard world model setting where the objective is to predict future observations given current state and action. Formally, we denote observations (e.g., video frames) as $\boldsymbol{O} = [\boldsymbol{o}_0, \boldsymbol{o}_1, \dots, \boldsymbol{o}_T] \in \mathbb{R}^{T\times H\times W\times 3}$ and actions as $\boldsymbol{A} = [\boldsymbol{a}_0, \boldsymbol{a}_1, \dots, \boldsymbol{a}_T] \in \mathbb{R}^{T\times 3}$[1]. The world model f_θ can be formulated as:

$$\begin{aligned} f_\theta : \mathbb{R}^{H\times W\times 3} \times \mathbb{R}^3 \to \mathcal{P}(\mathbb{R}^{H\times W\times 3}), \\ f_\theta(\boldsymbol{o}_t, \boldsymbol{a}_t) \mapsto p_\theta(\boldsymbol{o}_{t+1} | \boldsymbol{o}_t, \boldsymbol{a}_t). \end{aligned} \tag{1}$$

Because real-world dynamics are inherently uncertain and only partially observable, a world model should produce a stochastic distribution over future states rather than a single deterministic predictions.

Such stochastic formulation of the world model can be naturally implemented using generative modeling, where the generator is conditioned on past observation $\boldsymbol{o}_t$ and action $\boldsymbol{a}_t$. Direct generative modeling in pixel space is computationally prohibitive due to the high dimensionality of visual observations. Instead, the world model f_θ can be formulated to operate on low-dimensional latent tokens $\boldsymbol{z} \in \mathbb{R}^{N\times D}$ [3]. These latent tokens are obtained via an image tokenizer comprising an encoder $\mathcal{E} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{N\times D}$ and decoder $\mathcal{D} : \mathbb{R}^{N\times D} \to \mathbb{R}^{H\times W\times 3}$, trained with a reconstruction objective: $\mathcal{L}_{\text{recon}} = ||\boldsymbol{o} - \mathcal{D}(\mathcal{E}(\boldsymbol{o}))||_2^2$ (Fig. 1(a)). Extending Eq. 1, latent world model f_ϕ can be described as

$$\begin{aligned} f_\phi : \mathbb{R}^{N\times D} \times \mathbb{R}^3 \to \mathcal{P}(\mathbb{R}^{N\times D}), \\ f_\phi(\boldsymbol{z}_t, \boldsymbol{a}_t) \mapsto p_\phi(\boldsymbol{z}_{t+1} | \boldsymbol{z}_t, \boldsymbol{a}_t), \end{aligned} \tag{2}$$

where $\boldsymbol{z}_t = \mathcal{E}(\boldsymbol{o}_t)$. Here, the token count N directly determines computational complexity: for attention-based architectures [49] commonly used in generative models, cost scales quadratically with N. By keeping N small, the latent world model formulation alleviates this quadratic bottleneck and enables efficient decision-time planning.

Once the latent world model f_θ is trained, we can use it to find a sequence of actions $\{\boldsymbol{a}_t\}$ that drives the transition from the initial observation $\boldsymbol{o}_0$ to the goal observation $\boldsymbol{o}_{\text{goal}}$, as illustrated in Fig. 1(c). We first compute $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{o}_0)$, and initialize a candidate action sequence $\mathbf{a} = [\boldsymbol{a}_0, \boldsymbol{a}_1, \dots, \boldsymbol{a}_{H-1}]$. Then, we obtain a sequence of latent tokens $\{\boldsymbol{z}_t\}$ by rolling out the trained world model to predict future states over the planning horizon H:

$$\boldsymbol{z}_{t+1} \sim f_\phi(\boldsymbol{z}_t, \boldsymbol{a}_t),\ t \in \{0, \cdots, H-1\}. \tag{3}$$

After the rollout reaches the planning horizon (*i.e.*, $\boldsymbol{z}_H$ is sampled), the candidate action sequence $\mathbf{a}$ is evaluated using a cost function that measures the distance between the final predicted observation and the goal: $C(\mathbf{a}) = d(\hat{\boldsymbol{o}}_H, \boldsymbol{o}_{\text{goal}})$, where $\hat{\boldsymbol{o}}_H = \mathcal{D}(\boldsymbol{z}_H)$, $\hat{\boldsymbol{o}}_{\text{goal}} = \mathcal{D}(\boldsymbol{z}_{\text{goal}})$, and $d(\cdot,\cdot)$ is a distance measure (*e.g.* LPIPS [32]). The optimal action sequence is then obtained via solving: $\mathbf{a}^* = \arg\min_{\mathbf{a}} C(\mathbf{a})$, where the optimization can be performed using sampling-based methods [11, 12, 61] or gradient descent.

2.2. CompACT tokenizer

The computation bottleneck in world model planning stems from the latent token count N: conventional tokenizers typically encode images with hundreds of tokens, which slows down their sampling during autoregressive rollout. We introduce CompACT, a compact tokenizer $\mathcal{D}_{\text{compact}} \circ \mathcal{E}_{\text{compact}}$ that encodes each image into just 16 discrete tokens and avoids iterative denoising by using a discrete latent space (Fig. 2). Despite this extreme compression, CompACT still preserves the sufficient information for effective planning (Sec. 3).

Compact discrete encoding Our tokenizer encoder $\mathcal{E}_{\text{compact}} : \mathbb{R}^{H\times W\times 3} \to \{1, \dots, K\}^{16}$ maps an input image $\boldsymbol{o}$ into a sequence of 16 discrete tokens $\boldsymbol{z}$, each selected from a vocabulary of size K. The encoder architecture consists of a vision transformer [14] followed by a quantization

[1] In navigation settings, actions are 3-dimensional, representing changes in x-axis, y-axis, and yaw. The formulation generalizes to different action dimensions.

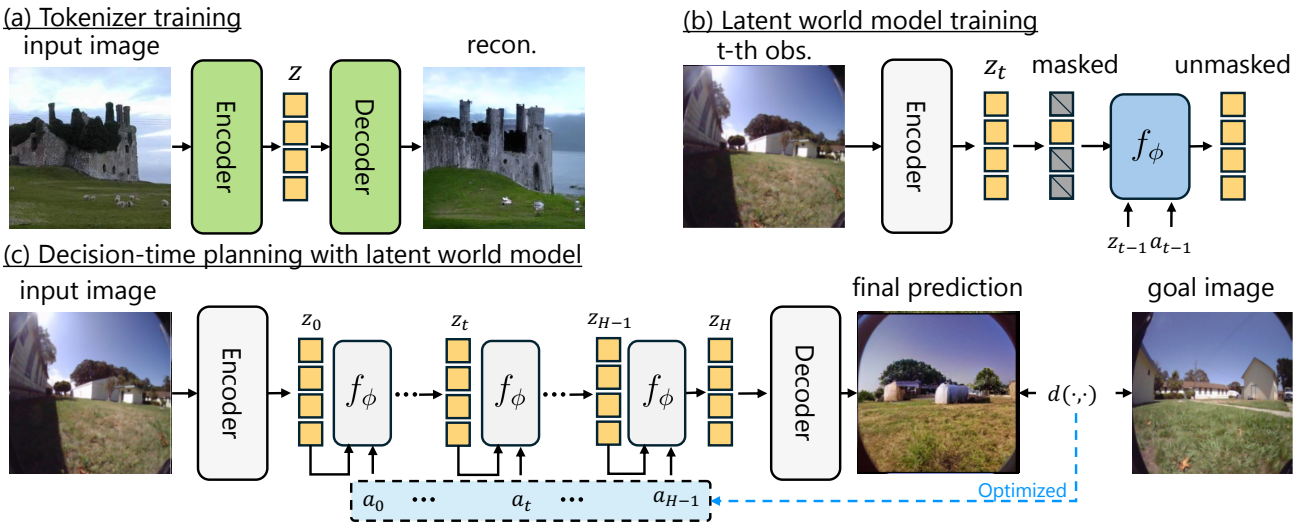

Figure 1. Overview of the proposed latent world model formulation (Sec. 2.1). (a) An image tokenizer is first trained with a reconstruction objective to map an input image into compact latent tokens $\boldsymbol{z}$. (Fig. 2 and Sec. 2.2). (b) Using the learned tokenizer, latent world model $f_\phi(\boldsymbol{z}_t, \boldsymbol{a}_t)$ is trained to model the conditional distribution of the future state $p_\phi(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \boldsymbol{a}_t)$, where we adopt masked generative modeling (Sec. 2.3). (c) At test time, the learned latent world model is used for *decision-time planning*: An optimization procedure (e.g., MPC with CEM) searches over actions $\boldsymbol{a}_{0:H-1}$ to minimize the distance between the predicted final state and a goal image.

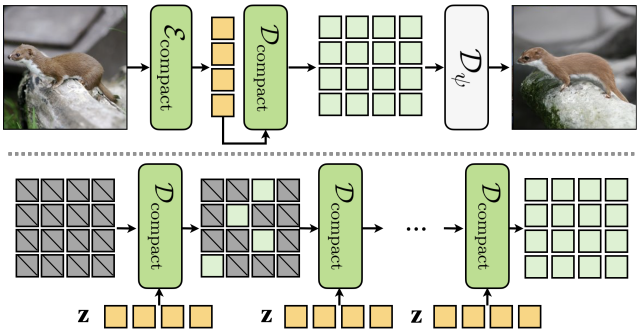

Figure 2. A figure for tokenizer architecture detail.

layer. Specifically, initial latent tokens $\boldsymbol{z}^0 \in \mathbb{R}^{16\times D}$ are concatenated with the image patch tokens and subsequently processed through a ViT. From the output of the ViT, only the tokens corresponding to the initial latent tokens are discretized using finite scalar quantization [43], yielding discrete latent tokens $\boldsymbol{z} \in \{1,\dots,K\}^{16}$. While extreme compression inevitably discards fine-grained visual details, we hypothesize that these details are largely irrelevant for planning tasks, where object-level semantics and spatial relationships dominate decision-making.

Generative decoding Direct pixel reconstruction from 16 tokens is an ill-posed problem—the information bottleneck prevents the deterministic recovery of high-frequency details, since diverse pixel-space manifestations can arise from identical low-frequency features. To address this, we propose a generative decoding strategy that sidesteps direct pixel reconstruction. Our decoder $\mathcal{D}_{\text{compact}} : \{1,\dots,K\}^{16} \to \{1,\dots,K_\psi\}^{N_\psi}$ learns to generate *latent tokens from a pretrained tokenizer* $\mathcal{D}_\psi \circ \mathcal{E}_\psi$ [39], using our compact tokens $\boldsymbol{z}$ as a condition. This transforms the intractable decompression problem into a conditional generation task. Specifically, we first convert an image $\boldsymbol{o}$ into target tokens $\boldsymbol{z}^\psi = \mathcal{E}_\psi(\boldsymbol{o}) \in \{1,\dots K_\psi\}^{N_\psi}$ using the pretrained tokenizer encoder, where $N_\psi \gg 16$ (typically $N_\psi = 196$ for 224×224 images). We then employ masked generative modeling [7, 64] to learn the mapping from $\boldsymbol{z}$ to $\boldsymbol{z}^\psi$, which offers significantly faster sampling than autoregressive [4, 55] models. During training, a random subset of the target tokens $\boldsymbol{z}^\psi$ is masked, and the decoder learns to recover them using the compact tokens $\boldsymbol{z}$ and the remaining unmasked tokens. The tokenizer training objective is defined to minimize the negative log-likelihood of the masked tokens $\boldsymbol{z}^\psi$:

$$\mathcal{L}_{\text{tok}} = -\mathbb{E}_{p(z^\psi)}\big[\log p(\boldsymbol{z}^\psi|\boldsymbol{z}, M(\boldsymbol{z}^\psi))\big], \tag{4}$$

where $M(\cdot)$ represents the random masking. During inference, $\mathcal{D}_{\text{compact}}$ begins with a fully masked sequence of a pretrained latent tokens and iteratively unmasks them following the sampling scheme based on its prediction confidence [7]. The compact tokens $\boldsymbol{z}$ provide high-level semantic guidance throughout this process, while the generative model synthesizes plausible visual details consistent with these semantics. The final reconstruction is obtained through the pretrained decoder: $\hat{\boldsymbol{o}} = (\mathcal{D}_\psi \circ \mathcal{D}_{\text{compact}} \circ \mathcal{E}_{\text{compact}})(\boldsymbol{o})$.

In a nutshell, our CompACT tokenizer achieves extreme compression by preserving only high-level semantics in 16 discrete tokens, then using these as conditioning for a generative decoder that synthesizes plausible high-frequency details. This design aligns with our core hypothesis that effective planning requires not photorealistic world models, but compact representations of decision-critical information.

2.3. Compact latent world model

With our CompACT tokenizer defined, we can now train the world model formulated in Eq. 2 directly in the 16-token discrete latent space, as described in Fig. 1(b). Given a dataset of observation and action sequence, we first encode all observation into compact latent tokens using CompACT tokenizer: $\boldsymbol{z}_t = \mathcal{E}_{\text{compact}}(\boldsymbol{o}_t)$. Similar to generative decoding, we use the masked generative modeling [7] to train the world model f_ϕ. The training objective is denoted as:

$$\mathcal{L}_{\text{world}} = -\mathbb{E}_{p(z_t, a_t, z_{t+1})}\big[\log p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \boldsymbol{a}_t, M(\boldsymbol{z}_{t+1}))\big]. \tag{5}$$

The key advantage of this formulation is computational efficiency during planning. During model-predictive control, it can now perform rollouts using only 16 tokens per timestep, enabling planning latency that was previously intractable with hundred-length tokens.

3. Experiment

3.1. Experimental Settings

To validate the effectiveness of CompACT tokenizer, we train a world model for navigation scenarios following NWM [3] and evaluate planning performance in model-predictive control settings. Due to the space constraints, we explain the details of the model architecture and the evaluation protocol in the supplementary material.

Dataset The CompACT is trained on ImageNet-1K [13], using the VQGAN from MAGE [39] as the pretrained tokenizer. The world model is trained on RECON [51], SCAND [33], and HuRoN [29], following NWM.

Tokenizer baselines We compare our approach against two baseline tokenizers: (1) **SD-VAE** [50]: A continuous latent space tokenizer used in NWM that requires 196 tokens to encode a 224×224 image. (2) **FlexTok** [2]: A recently proposed tokenizer with discrete latent space that enables dynamic truncation of latent token sequences. Early tokens encode semantic information and low-frequency features, while later tokens capture high-frequency visual details. We used first 16 and 64 tokens for coomparison. For the world model, we maintain identical architectures (200M parameters) and hyperparameters across all baselines, except that discrete latent tokens require an additional linear layer to predict token logits.

Tokenizer	#tok	type	rFID↓	APE	RPE	Latency
SD-VAE [50]	196	cont.	0.98	**1.262**	**0.354**	145.0
FlexTok [2]	64	disc.	4.18	1.484	0.400	14.5
	16	disc.	4.26	1.625	0.446	<u>13.6</u>
CompACT (Ours)	16	disc.	7.28	<u>1.330</u>	<u>0.390</u>	**7.4**

Table 1. **Planning performance of NWM on RECON benchmark with different tokenizers.** rFID (reconstruction FID) measures reconstruction quality on ImageNet [13] validation split. Latency (sec) represents single trajectory optimization time using 4 A6000 ADA GPUs.

3.2. Experimental results

Planning performance Table 1 presents planning results for goal-conditioned visual navigation on the RECON dataset. Our CompACT achieves a 20× reduction in planning latency while maintaining comparable planning accuracy. Interestingly, the NWM trained with our tokenizer outperforms the FlexTok-based model at both 16 and 64 token configurations. We attribute this to FlexTok's random token truncation during training, which ablates spatial information from the earlier tokens such as object layout that is critical for planning. In contrast, our tokenizer—trained without truncation—must encode all necessary information within exactly 16 tokens, ensuring that planning-relevant features are consistently preserved in this compact representation.

Reconstruction performance Table 1 presents the reconstruction quality of each tokenizer measured by reconstruction FID (rFID). The results reveal that high reconstruction fidelity does not translate to better downstream planning performance. This finding suggests that decision-time planning can be made significantly more efficient by employing extreme compression tokenizers like CompACT, which prioritize planning-relevant features over pixel-level reconstruction quality.

4. Conclusion and Future Work

In this work, we present CompACT, a compact tokenizer that achieves extreme compression by representing images with only 16 discrete tokens. We demonstrate that this approach enables highly efficient world models, outperforming baselines requiring larger token counts while achieving a 20× speedup by preserving only planning-critical features.

For future work, we identify two promising directions: (1) We are developing methods to explicitly enforce the tokenizer to discover planning-critical features within observations, which we believe will further improve downstream planning performance. (2) We aim to address the lack of proper metrics for measuring the *plannability* of learned representations, as this remains a key bottleneck in evaluating and optimizing a tokenizer for the world model.

Acknowledgments

This work was supported by the IITP grants (RS-2024-00457882, RS-2025-25443730) funded by Ministry of Science and ICT, Korea.

References

[1] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37: 58757–58791, 2024. 1, 2

[2] Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *Forty-second International Conference on Machine Learning*, 2025. 2, 4, 1

[3] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025. 1, 2, 4

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 3

[5] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 2

[6] Shiyue Cao, Yueqin Yin, Lianghua Huang, Yu Liu, Xin Zhao, Deli Zhao, and Kaigi Huang. Efficient-vqgan: Towards high-resolution image generation with efficient vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7368–7377, 2023. 1

[7] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 1, 3, 4, 2

[8] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *ICML*, 2023. 1

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 1

[10] Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. *arXiv preprint arXiv:1903.01959*, 2019. 2

[11] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018. 2, 1

[12] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005. 1, 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 1

[15] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 1

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[17] Cunxin Fan, Xiaosong Jia, Yihang Sun, Yixiao Wang, Jianglan Wei, Ziyang Gong, Xiangyu Zhao, Masayoshi Tomizuka, Xue Yang, Junchi Yan, et al. Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions. *arXiv preprint arXiv:2505.02152*, 2025. 2

[18] Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024. 1

[19] Jay W Forrester. Counterintuitive behavior of social systems. *Theory and decision*, 2(2):109–140, 1971. 1

[20] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 1

[21] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024. 2

[22] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[23] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018. 1, 2

[24] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. 1, 2

[25] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. 2

[26] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 1, 2

[27] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023. 1

[28] Noriaki Hirose, Fei Xia, Roberto Martín-Martín, Amir Sadeghian, and Silvio Savarese. Deep visual mpc-policy learning for navigation. *IEEE Robotics and Automation Letters*, 4(4):3184–3191, 2019. 2

[29] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023. 4

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1

[31] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2

[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[33] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7 (4):11807–11814, 2022. 4

[34] Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. *arXiv preprint arXiv:2501.07730*, 2025. 1

[35] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14738–14748, 2021. 2

[36] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11523–11532, 2022. 1

[37] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and WOOK SHIN HAN. Draft-and-revise: Effective image generation with contextual rq-transformer. *Advances in Neural Information Processing Systems*, 35:30127–30138, 2022. 1

[38] José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. In *European Conference on Computer Vision*, pages 70–86. Springer, 2022. 1

[39] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *CVPR*, 2023. 3, 4, 1

[40] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 2024. 1

[41] Travis Manderson, Juan Camilo Gamboa Higuera, Stefan Wapnick, Jean-François Tremblay, Florian Shkurti, David Meger, and Gregory Dudek. Vision-based goal-conditioned policies for underwater navigation in the presence of obstacles. *arXiv preprint arXiv:2006.16235*, 2020. 2

[42] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*, 2023. 2

[43] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023. 3, 1

[44] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022. 1

[45] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016. 2

[46] Keita Miwa, Kento Sasaki, Hidehisa Arai, Tsubasa Takahashi, and Yu Yamaguchi. One-d-piece: Image tokenizer meets quality-controllable compression. *arXiv preprint arXiv:2501.10064*, 2025. 1

[47] Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. Wmnav: Integrating vision-language models into world models for object goal navigation. *arXiv preprint arXiv:2503.02247*, 2025. 2

[48] Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Yide Shentu, Evan Shelhamer, Jitendra Malik, Alexei A Efros, and Trevor Darrell. Zero-shot visual imitation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2050–2053, 2018. 2

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1, 2

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 4

[51] Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. In *5th Annual Conference on Robot Learning*, 2021. 2, 4

[52] Dhruv Shah, Arjun Bhorkar, Hrish Leen, Ilya Kostrikov, Nick Rhinehart, and Sergey Levine. Offline reinforcement learning for visual navigation. *arXiv preprint arXiv:2212.08244*, 2022. 2

[53] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023. 2

[54] Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 63–70. IEEE, 2024. 2

[55] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 3

[56] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 1

[57] Stephen Tian, Chelsea Finn, and Jiajun Wu. A control-centric benchmark for video prediction. *arXiv preprint arXiv:2304.13723*, 2023. 1

[58] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 2

[59] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[60] Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. Maskbit: Embedding-free image generation via bit tokens. *arXiv preprint arXiv:2409.16211*, 2024. 1

[61] Grady Williams, Paul Drews, Brian Goldfain, James M Rehg, and Evangelos A Theodorou. Aggressive driving with model predictive path integral control. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 1433–1440. IEEE, 2016. 1, 2

[62] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023. 1, 2

[63] Xuan Yao, Junyu Gao, and Changsheng Xu. Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. *arXiv preprint arXiv:2506.23468*, 2025. 2

[64] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 3, 1

[65] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024. 1

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1

[67] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10412–10420, 2025. 2

[68] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022. 1

[69] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. 1

[70] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024. 1

Planning in 16 Tokens: A Compact Discrete Tokenizer for Latent World Model

Supplementary Material

5. Experiment Details

Model architecture For the compact tokenizer, we employ ViT [14] and DiT [49] as the encoder $\mathcal{E}_{\text{compact}}$ and decoder $\mathcal{D}_{\text{compact}}$, respectively. For the world model, we adopt CDiT, a DiT variant proposed in NWM.

Evaluation. We evaluate planned trajectories using two metrics: absolute trajectory error (ATE) and relative pose error (RPE). For planning, we optimize action sequences using the cross-entropy method [11, 12] with 80 candidate action sequences. All other hyperparameters for the world model follow the configuration in NWM.

6. Qualitative results of planning

Fig. 3 presents an example planning result with the proposed CompACT. While the generated simulations lose fine-grained details such as textures or shadows, they preserve planning-critical features such as the overall scene layout.

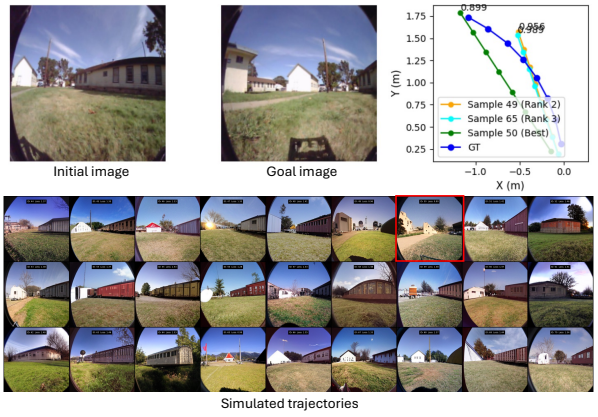

Figure 3. Qualitative results of planning with the proposed CompACT. The final rollout corresponding to the simulated trajectory with the minimum cost is highlighted in red.

7. Related Work

7.1. Image tokenization

Image tokenization has played a crucial role in visual generation. Encoding raw image pixels into compressed latent representations alleviates the difficulty of directly modeling distributions in high-dimensional continuous spaces. By discretizing these latents, the model can efficiently produce and sample categorical distributions for individual tokens. Autoencoder-based architectures such as VQ-VAE [59], VQGAN [16], ViT-VQGAN [50], and Efficient-VQGAN [6] employ vector quantization to form discrete latent spaces. Numerous enhancements have been proposed to improve reconstruction quality, including perceptual losses [66], adversarial losses [22], transformer-based designs, residual quantization [36], lookup-free quantization [64], and finite scalar quantization (FSQ) [43].

A common limitation of the aforementioned approaches is their reliance on 2D patch-grid latent representations. This design fixes the number of tokens according to the image resolution (H, W) and prevents its adaptive adjustment based on image complexity. To overcome this, recent works have explored 1D tokenization [2, 34, 46, 65], which does not explicitly preserve spatial structure. For example, TiTok [65] employs learned register tokens to capture image content in a compact sequence via a ViT encoder, enabling efficient representation learning. Although TiTok [65] achieves highly compact tokenization, it encodes each image using a fixed set of 32 tokens, regardless of the image's complexity. FlexTok [2] addresses these constraints by allowing flexible token lengths ranging from 1 to 256 tokens and employing a coarse-to-fine design in which later tokens capture progressively finer details.

As such, existing image tokenizers prioritize high-frequency details, which are often unnecessary for downstream planning tasks. For this reason, we believe a planning-oriented tokenizer would be more effective for applications like robotic navigation. By using a smaller number of tokens to represent a scene, our proposed 1D tokenizer allows agents to simulate more scenarios and find optimal plans faster, aggressively compressing token length while preserving only the essential visual information.

7.2. Masked Autoregressive Image Generation

Masked autoregressive image generation models [7, 8, 18, 20, 39, 40, 60] leverage encoder–decoder architectures with bidirectional attention mechanisms to reconstruct masked tokens during generation. Unlike traditional autoregressive (decoder-only) models [9, 56], which predict tokens sequentially, these architectures [64, 68] perform parallel decoding. They predict multiple token positions in a single step, thereby reducing the number of steps needed for full image generation while improving inference efficiency. Notably, MaskGIT [7] and MAR [40] have demonstrated that such designs enable both rapid and high-quality image synthesis. In parallel, research on advanced sampling strategies has emerged [37, 38], aiming to further improve generation quality and convergence speed.

In this work, we focus on the tokenization stage and adopt the widely used non-autoregressive sampling approach from MaskGIT [7] for generating token sequences that are subsequently decoded into 2D discrete latent tokens that serve as the input to the VQ-GAN decoder in our pipeline.

7.3. Goal Conditioned Visual Navigation

Goal-conditioned visual navigation [35, 45, 48, 53, 54] constitutes a key challenge in robotics, as it requires the integration of both perception and planning capabilities. Given one or more context images along with a navigation goal image, such models [53, 54] aim to produce an optimal path to the goal when the environment is known, or to explore the surroundings otherwise.

Previously, this task has been addressed through policy-based approaches [10, 28, 41, 52–54], in which agents learn a direct mapping from observations to actions. Methods in this category often employ reinforcement learning, behavior cloning, or model-free exploration strategies, and are typically optimized for specific environments or goal conditions.

Recent studies have shifted toward leveraging world models[23], which aim to capture and distill knowledge of complex, high-dimensional environments. An agent with a world model can predict its future by mentally simulating the outcomes of a sequence of proposed actions. Consequently, it enables the agent to perform look-ahead reasoning before acting. Rather than directly outputting actions, world models simulate the environment by taking the current state or observation and the policy's action as input and predicting the next latent state along with an associated reward. The Dreamer series[24–26] exemplifies this approach by modeling environmental dynamics that facilitate long-horizon planning and control.

Beyond simulated control domains, world models have been successfully used in robotics [17, 42, 62], games [1, 5, 24, 25, 58], autonomous driving [21, 31, 67] and navigation [3, 35, 47, 63]. Specifically in visual navigation, Pathdreamer [35] generates high-resolution RGB, depth, and semantic panoramas of future viewpoints from panoramic indoor observations. Most recently, NWM[3] introduced a Conditional Diffusion Transformer to simulate trajectories for planning, eliminating the need for explicit 3D reconstructions or geometric priors while maintaining generality across diverse environments. However, NWM[3] requires 196 tokens by the SD-VAE tokenizer. Differently, we use a compact image tokenizer that encodes the scene into a much smaller number of tokens—just 16 tokens—representing an extreme compression, and leverage a Multimodal Diffusion Transformer as the generative decoder.