PredToken: Predicting Unknown Tokens and Beyond with Coarse-to-Fine Iterative Decoding

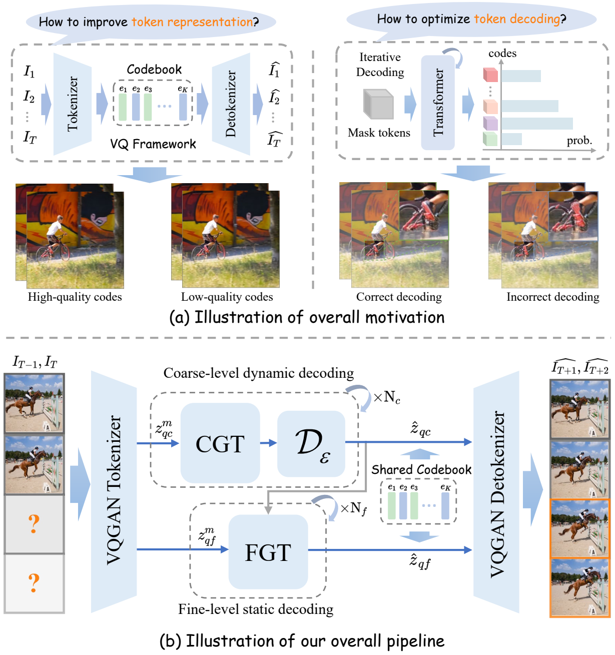
Xuesong Nie[1*] Haoyuan Jin[1*] Yunfeng Yan[1†] Xi Chen[2†] Zhihang Zhu[1] Donglian Qi[1]
[1]Zhejiang University [2]The University of Hong Kong
xuesongnie@zju.edu.cn

Abstract

Predictive learning models, which aim to predict future frames based on past observations, are crucial to constructing world models. These models need to maintain low-level consistency and capture high-level dynamics in unannotated spatiotemporal data. Transitioning from frame-wise to token-wise prediction presents a viable strategy for addressing these needs. How to improve token representation and optimize token decoding presents significant challenges. This paper introduces PredToken, a novel predictive framework that addresses these issues by decoupling space-time tokens into distinct components for iterative cascaded decoding. Concretely, we first design a "decomposition, quantization, and reconstruction" schema based on VQGAN to improve the token representation. This scheme disentangles low- and high-frequency representations and employs a dimension-aware quantization model, allowing more low-level details to be preserved. Building on this, we present a "coarse-to-fine iterative decoding" method. It leverages dynamic soft decoding to refine coarse tokens and static soft decoding for fine tokens, enabling more high-level dynamics to be captured. These designs make PredToken produce high-quality predictions. Extensive experiments demonstrate the superiority of our method on various real-world spatiotemporal predictive benchmarks. Furthermore, PredToken can also be extended to other visual generative tasks to yield realistic outcomes.

1. Introduction

With human cognition evolving, granting us greater foresight, equipping machines with this foresight remains a significant challenge due to the world's inherent chaos. Predictive learning, an unsupervised method focused on predicting future events based on past observations, plays a critical role in the world models [11, 17, 47]. This approach, unlike supervised models that depend on annotated data, leverages unannotated data to autonomously uncover complex spatial and temporal patterns, holding the promise to revolutionize domains where labeled data is scarce [16, 22, 33, 34].



(a) Illustration of overall motivation

(b) Illustration of our overall pipeline

Figure 1. (a) The illustration of our motivation highlights the critical importance of improving token representation and decoding mechanisms for high-quality predictive learning. (b) The illustration of our pipeline includes the DQR-based VQ framework and the coarse-to-fine iterative decoding method, ensuring low-level consistency and high-level dynamics in spatiotemporal data.

Benefiting from the token-wise learning paradigm, recent endeavors have provided valuable insight into predictive learning. They attempt to simplify the issue by decomposing frame-level prediction into token-level prediction. Walker et al. [38] leverage a vector quantized variational autoencoder [37] (VQVAE) for compressing video into discrete latent tokens, subsequently decoding via a prior model PixelCNN [36] with an autoregressive manner. Gupta et al.

[*] Equal Contribution.
[†] Corresponding Author.

[15] further enhances the discrete token representation by utilizing VQGAN [8] and coupled with a bi-directional window Transformer for iterative token decoding. This parallel decoding approach overcomes the time-consuming drawback of traditional autoregressive algorithms. Yu et al. [49] advances this technique to diverse video generation tasks by incorporating a 3D version of VQGAN and a novel token masking strategy. These approaches typically involve training multiple independent and large models to tackle visual tasks collaboratively. Specifically, their general pipeline often comprises two stages: (i) leveraging vector quantization framework to represent visual data into discrete latent codes during stage 1, and (ii) modeling the data distribution via a prior model (*e.g*., CNN or ViT) in discrete space during stage 2. Nonetheless, current architectures still struggle to maintain low-level consistency and capture high-level dynamics in intricate environments. We will explore this framework from the following two aspects.

Firstly, in stage 1, learning discrete token representation is crucial for restoring low-level spatiotemporal details. Figure. 1(a) left shows that high-quality codes retain more visual information than low-quality ones. Some methods [12, 39] employ vector quantization techniques to spatiotemporal data with vanilla VQGAN or 3D VQGAN. Meanwhile, other approaches [3, 48] enhance token representation by incorporating spatial attention mechanisms. These efforts indicate that both improving the quantization framework and refining the network architecture are effective. Increasing resolution enriches visual detail in frames, where prediction errors often emerge around high-frequency features. This observation prompts the question of whether a network can be trained to capture both low and high-frequency details accurately. Therefore, we designed a “decomposition, quantization, and reconstruction” (DQR) schema coupled with a dimension-aware quantization model to better preserve low-level visual details.

Secondly, in stage 2, the token decoding schema plays a key role in capturing high-level spatiotemporal dynamics. Figure 1(a) right illustrates a simple decoding example where correct decoding benefits adjacent tokens. In contrast, incorrect decoding negatively impacts nearby tokens in the spatiotemporal context, hindering the learning of complex motion patterns. Current frameworks utilize either autoregressive or non-autoregressive iterative methods for decoding tokens. However, a decoding mistake in these methods can negatively impact the decoding of nearby spatiotemporal tokens. The vanilla iterative decoding suffers from two problems: (i) the spatiotemporal decoding process is irreversible, indicating that incorrect tokens cannot be updated in later iterations. (ii) the sampling process does not consider the correlation between tokens within the codebook, *e.g*., tokens describing the same moving object should interact with each other. Therefore, we introduce a mask discriminator and soft decoding to improve the token iterative decoding process.

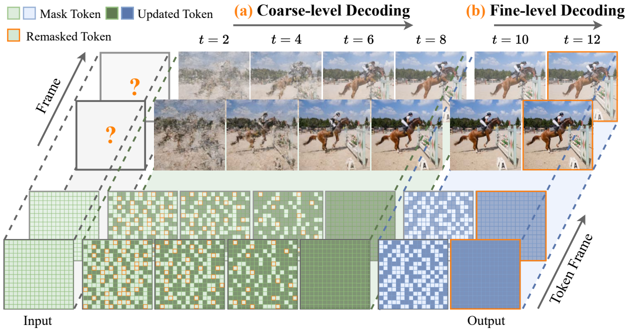


Figure 2. The visualization of the coarse-to-fine iterative decoding process. Green for coarse tokens, blue for fine tokens, and orange box for tokens dynamically remasked by a mask discriminator.

In this paper, we present PredToken, an innovative predictive approach that highlights the significance of both improving token representation and optimizing token decoding for superior predictions. First, we propose a “decomposition, quantization, and reconstruction” schema based on VQGAN to enhance the token representation, which disentangles low- and high-frequency details via wavelet transform and incorporates a dimension-aware quantization model to retain more low-level information. Building on this, we then propose a “coarse-to-fine iterative decoding” strategy that guides the unknown or corrupted tokens to accurate tokens. This strategy employs dynamic soft decoding for coarse token refinement, which then cascades into static soft decoding for fine token enhancement. Benefiting from token-level spatiotemporal data modeling, PredToken alleviates the burden of handling millions of pixels. These designs enable PredToken to not only excel in spatiotemporal predictive benchmarks but also produce realistic outcomes when extended to other visual generative tasks.

2. Related Work

Spatiotemporal Predictive Learning. Over time, many recurrent-based models have provided significant insights for predictive learning. ConvLSTM [28] seamlessly incorporates 2D convolution into the recurrent state transitions of standard LSTM, proposing the convolutional LSTM unit. PredRNN [40] further enhances convolutional LSTMs with pairwise memory cells to capture long-term and short-term patterns. Conv-TT-LSTM [32] introduces a higher-order convolutional LSTM model with a novel convolutional tensor-train decomposition for long-term prediction. E3D-LSTM [42] extends LSTM with 3D convolution. SwinLSTM [35] integrates the Swin Transformer [21] module into LSTM for improving performance. In addition to the recurrent-based models, recent literature [22, 45, 53] attempts to build recurrent-free models that predict future sequences in parallel for efficiency. SimVP [10] utilizes Inception modules with a UNet architecture to learn the temporal evolution. TAU [33] proposes a temporal attention unit to capture long-term temporal dependencies. DMVFN [18] proposes a dynamic voxel flow network for video prediction. Unlike the above models, PredToken focuses on maintaining low-level consistency and capturing high-level dynamics in spatiotemporal data.

Vector-Quantized Generative Models. Inspired by GPT [2], many pioneering works tokenize various types of data (*e.g.*, images, audio, and videos) into discrete tokens via vector quantization and train a prior model to generate content. In the image domain, VQVAE [37] converts images into discrete tokens and models their patterns with an autoregressive model. VQGAN [8] enhances image fidelity by introducing adversarial training and perceptual loss. MaskGIT [4] introduces a novel non-autoregressive generation paradigm with masked token modeling. Token-Critic [20] improves the sampling quality of pretrained generative models by training an additional network. StraIT [24] proposes image stratification that obtains an interlinked token pair to improve generation capabilities. Vision ELECTRA [52] introduces adversarial masked image modeling with a hierarchical discriminator to improve reconstruction. CIM [9] employs an auxiliary generator and an enhancer for image corruption and reconstruction for self-supervised pretraining. In the video domain, Video VQ-VAE [38] compresses videos into hierarchical discrete tokens and predicts future frames with PixelCNN [36] models. VideoGPT [46] represents videos as tokens and generates them with GPT [2] models. MaskViT [15] combines 2D VQGAN with a bidirectional window Transformer for frame prediction. MAGVIT [49] introduces a 3D VQGAN and a conditional masked modeling strategy for multi-task video generation. WorldDreamer [39] employs multimodal tokenizers and a spatial-temporal patchwise Transformer for video generation. PredToken improves existing frameworks by enhancing the token representation and optimizing the token decoding for high-quality spatiotemporal prediction.

3. Proposed Method

The PredToken model comprises three stages: two training stages and one inference stage. Stage I (Sec. 3.1) utilizes a VQGAN-based "decomposition, quantization, reconstruction" framework for vector quantization learning. Stage II (Sec. 3.2) involves training two prior models (coarse-grained and fine-grained Transformers) and a mask discriminator. Stage III (Sec. 3.3) adopts a "coarse-to-fine iterative decoding" strategy for better token decoding. The details of each stage will be explained below.

3.1. Stage I: Learning Vector Quantization

We design a VQGAN-based "decomposition, quantization, reconstruction" framework to improve the token representation by disentangling frequency components, as shown in Figure Stage I. Our approach involves three steps:

Step1: Decomposition. We introduce the discrete wavelet transform (DWT) to improve the vector quantization framework. For signal t, given the wavelet function $\psi_{j,k}(t) = 2^{\frac{j}{2}}\psi\left(2^j t - k\right)$ with scaling factor j and time factor k, and scale function $\phi_{j,k}(t) = 2^{\frac{j}{2}}\phi\left(2^j t - k\right)$, the decomposition at level j_0 is given by:

$$f(t) = \sum_{j>j_0}\sum_{k} d_{j,k}\psi_{j,k}(t) + \sum_{k} c_{j_0,k}\phi_{j_0,k}(t), \tag{1}$$

where $d_{j,k} = \langle f(t), \psi_{j,k}(t)\rangle$ and $c_{j_0,k} = \langle f(t), \phi_{j_0,k}(t)\rangle$ represent the detail and approximation coefficients, respectively. The high-pass and low-pass filter are denoted with $\vec{a}_0[k] = \left\langle \frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right), \phi(t-k)\right\rangle$ and $\vec{a}_1[k] = \left\langle \frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right), \phi(t-k)\right\rangle$. We extend it to 2D filters by vector outer products, which can be formulated as:

$$\mathcal{F}_{\mathrm{LL}} = \vec{a_0} \times \vec{a_0}^T,\ \mathcal{F}_{\mathrm{LH}} = \vec{a_0} \times \vec{a_1}^T, \tag{2}$$

$$\mathcal{F}_{\mathrm{HL}} = \vec{a_1} \times \vec{a_0}^T,\ \mathcal{F}_{\mathrm{HH}} = \vec{a_1} \times \vec{a_1}^T, \tag{3}$$

where the low-frequency filter $\mathcal{F}_{\mathrm{LL}}$ learns coarse-grained tokens, while the combination of horizontal, vertical, and diagonal high-frequency filters $\mathrm{Cat}(\mathcal{F}_{\mathrm{LH}}, \mathcal{F}_{\mathrm{HL}}, \mathcal{F}_{\mathrm{HH}})$ learns fine-grained tokens.

Step2: Quantization. Previous works [4, 49] employ a vision transformer with spatial global attention, overlooking modeling in other dimensions, such as motion between frames. To better model spatiotemporal data, we adopted a dimension-aware quantization model that explores self-attention mechanisms in time, space, and channel dimensions. We reduce computational overhead through computing space and channel attention in restricted windows and groups. For the encoded coarse and fine-grained embeddings $\hat{z}_c, \hat{z}_f \in \mathbb{R}^{T\times C\times H/f\times W/f}$, the vector quantization process is described as searching for the nearest code in the learnable codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K}$, where $z_k \in \mathbb{R}^C$ denotes the k-th discrete token, K and C is the codebook and dimension size, respectively. It can be formulated as:

$$z_q = \underset{z_k\in\mathcal{Z}}{\mathrm{argmin}}\,\|\hat{z} - z_k\|\,, \tag{4}$$

where the quantization process is non-differentiable due to argmin operator, we adopt the straight-through gradient estimator [8, 37], enabling end-to-end training using the codebook-learning loss function:

$$\begin{aligned}\mathcal{L}_{\mathrm{VQ}} &= \|\mathrm{sg}(\hat{z}_c) - z_{qc}\|_2^2 + \beta\cdot\|\hat{z}_c - \mathrm{sg}(z_{qc})\|_2^2 \\ &+ \|\mathrm{sg}(\hat{z}_f) - z_{qf}\|_2^2 + \beta\cdot\|\hat{z}_f - \mathrm{sg}(z_{qf})\|_2^2,\end{aligned} \tag{5}$$

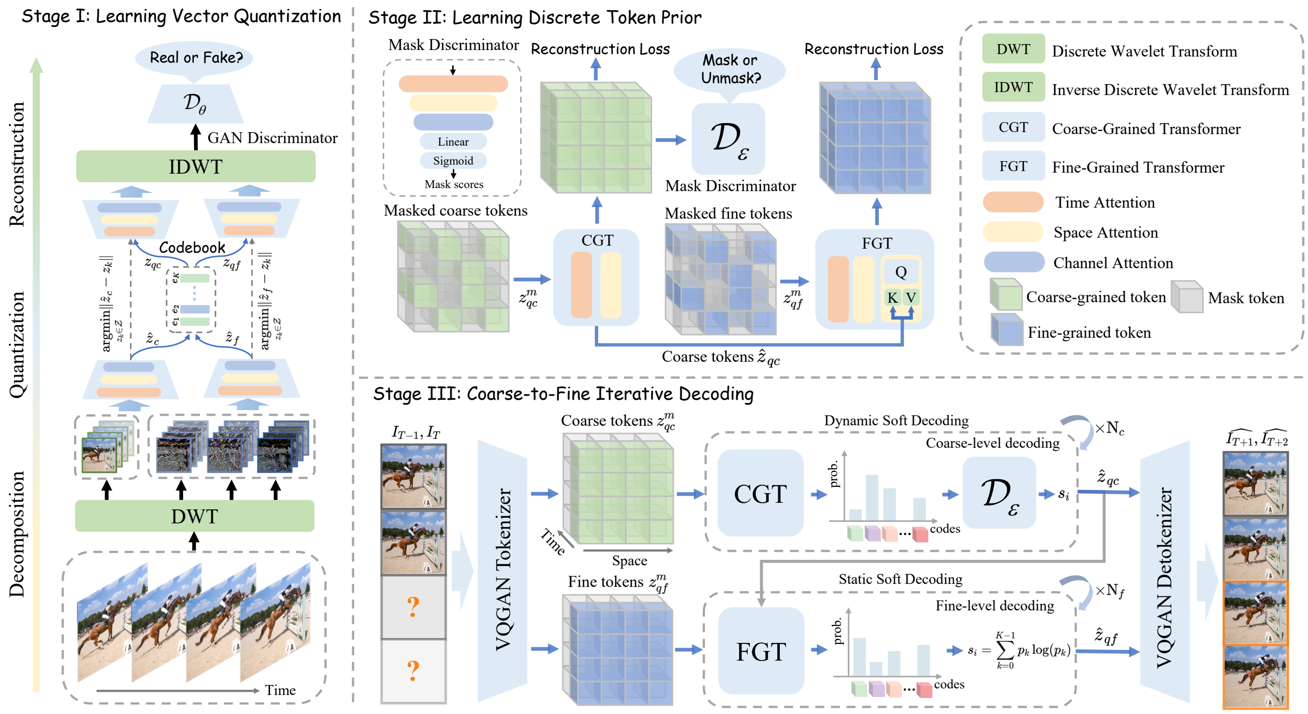


Figure 3. Overall architecture of PredToken. Stage I learns the discrete token representation through a "decomposition, quantization, and reconstruction" (DQR) schema coupled with a dimension-aware quantization model, capturing inter-frame dynamics and intra-frame static features by exploring self-attention mechanisms in time, space, and channel dimensions. In Stage II, CGT and FGT are trained by predicting special masked `[MASK]` tokens on spatiotemporal sequences, while the mask discriminator is trained through adversarial learning against CGT. Stage III involves decoding visual tokens via the "coarse-to-fine iterative decoding" method.

where $\mathrm{sg}(\cdot)$ represents the stop-gradient operator and $\beta = 0.25$ is a weighting hyper-parameter to control the update frequency of the codebook $\mathcal{Z}$.

Step3: Reconstruction. The quantized coarse and fine-grained embeddings $z_{qc}, z_{qf} \in \mathbb{R}^{T\times C\times H/f\times W/f}$ further are decoded via the decoder and inverse discrete wavelet transform (IDWT) for detailed reconstruction. We introduce pixel-level $\mathcal{L}_2$ loss and a perceptual-level $\mathcal{L}_\mathrm{P}$ loss that utilizes pretrained VGG [29] features to stabilize the codebook-learning loss $\mathcal{L}_\mathrm{VQ}$. To mitigate training instability due to increased resolution, the projected GANs discriminator [26] $\mathcal{D}_\theta$ is calculated to produce adversarial loss $\mathcal{L}_\mathrm{GAN}$. The total training loss is defined as follows:

$$\mathcal{L} = \min_{E,D,\mathcal{Z}} \left(\max_{\mathcal{D}_\theta} \left(\lambda \mathcal{L}_\mathrm{GAN} \right) + \mathcal{L}_2 + \mathcal{L}_\mathrm{P} + \mathcal{L}_\mathrm{VQ} \right), \quad (6)$$

where λ is adaptive weight [8], E and D denote encoder and decoder.

3.2. Stage II: Learning Discrete Token Prior

The learning of discrete tokens in prior models employs either autoregressive or non-autoregressive methods, where the former predicts the next token while the latter predicts special masked `[MASK]` tokens via the masked token modeling (MTM) task. In PredToken, the trainable network comprises a coarse-grained Transformer (CGT), fine-grained Transformer (FGT), and mask discriminator $\mathcal{D}_\varepsilon$. Both CGT and FGT are trained using MTM on spatiotemporal sequences, while the mask discriminator $\mathcal{D}_\varepsilon$ is trained through adversarial learning against CGT. Notably, adversarial training is conducted solely at the coarse level, as this level predominantly captures spatiotemporal information.

Given the discrete token sequences z_{qc}, z_{qf} from VQGAN tokenizer, as illustrated in Figure Stage II, the masked sequences z_{qc}^m, z_{qf}^m are created by uniformly sampling and replacing $\lceil \gamma(u) \cdot N \rceil$ tokens in z_{qc}, z_{qf} with the `[MASK]` token. The number of masked tokens is determined by the masking scheduler $\gamma(r) = \mathrm{cosine}(u\pi/2) \in (0, 1]$, where u is a scalar from 0 to 1. The CGT and FGT aim to reconstruct the original token sequences z_{qc}, z_{qf} based on those visible within the corrupted token sequences z_{qc}^m, z_{qf}^m. The objective of coarse- and fine-level reconstruction is to minimize the cross-entropy loss between the predicted and the original tokens at each masked position:

$$\mathcal{L}_\mathrm{Rec} = -\mathbb{E}\left[\sum_{i\in c,f} \sum_{z_{ij}^m = \texttt{[MASK]}} \log p_i(z_{ij} \mid z_i^m) \right], \quad (7)$$

where omitting subscript q for simplicity. The mask discriminator $\mathcal{D}_\varepsilon$ aims to differentiate whether the corresponding token is masked or unmasked. It is optimized using an adversarial training procedure:

$$\mathcal{L}_{\mathrm{GAN}} = \mathbb{E}\left[\log \mathcal{D}_\varepsilon(z_c) + \log\left(1 - \mathcal{D}_\varepsilon(\hat{z}_c)\right)\right], \tag{8}$$

where $\hat{z}_c$ denotes the reconstruction output of CGF. The total training loss can be formulated as:

$$\mathcal{L} = \min_{C,F}(\max_{\mathcal{D}_\varepsilon}(\alpha\mathcal{L}_{\mathrm{GAN}}) + \mathcal{L}_{\mathrm{Rec}}), \tag{9}$$

where α controls the relative importance. C and F represent CGT and FGT, respectively.

3.3. Stage III: Coarse-to-Fine Iterative Decoding

During inference, we employ models from stage I and II for the "coarse-to-fine iterative decoding" algorithm. This includes two levels: coarse-level dynamic soft decoding (DSD) and fine-level static soft decoding (SSD). For clarity, we will first detail the static soft decoding.

Static Soft Decoding. The discrete tokens from vector quantization are interrelated and should be considered collectively. Unlike previous methods [4, 49] that sample tokens from a categorical distribution, we employ soft sampling for iterative decoding, as shown in Figure 4. This method computes a weighted average of tokens based on predicted probabilities, thus preserving token correlations. Moreover, using probability values as scores is suboptimal because it neglects the overall distribution. Instead, we utilize the negative information entropy of the predicted distributions as a score to decide which tokens to retain or update. In the t-th iteration, SSD follows three steps:

(i) Non-Autoregressive Prediction. Given the masked sequences z_i^m at the current iteration, the model predicts the probabilities $p_i \in \mathbb{R}^{B\times T\times N\times K}$ in parallel.

(ii) Soft Probability Sampling. For each masked location i, we compute soft embeddings e_i by weighted average e_k according to the probability distribution, with corresponding scores s_i calculated using negative information entropy:

$$e_i = \sum_{k=0}^{K-1} p_k e_k, \quad s_i = \sum_{k=0}^{K-1} p_k \log(p_k). \tag{10}$$

where e_k denotes the k-th embedding in the codebook. This concept can be utilized in the VQGAN detokenizer. The scores of the masked tokens range from $-\log_2(K)$ to 0, while unmasked tokens receive positive scores.

(iii) Mask Scheduling and Updating. Based on the mask scheduling function γ, determine the number of tokens to mask or update $n = \lceil\gamma(\frac{t}{T})N\rceil$ and retain $\overline{n} = (N - n)$ each iteration. Following score-based sorting, the last n tokens are masked for updating in the next iteration.

Dynamic Soft Decoding. The static iterative decoding relies on the probability distribution of prior models, which

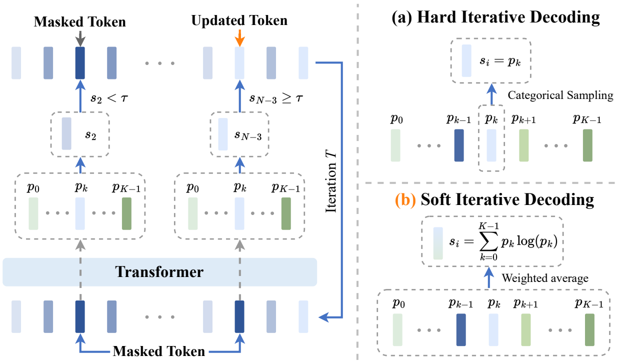


Figure 4. Overview of two static iterative decoding methods: (a) hard iterative decoding and (b) soft iterative decoding. The threshold τ is obtained as the $\lceil\gamma(\frac{t}{T})N\rceil$-th smallest score after sorting.

makes updating each token independent and static, meaning that tokens sampled incorrectly will not be corrected in subsequent iterations. The mask discriminator $\mathcal{D}_\varepsilon$ obtained in stage II elegantly addresses this issue by outputting a score between 0 and 1 for each token, dynamically correcting any tokens that were decoded incorrectly, as shown in Figure 2(a). Unlike the GAN discriminator [14, 19], used only during training, the mask discriminator not only boosts performance during training but also serves as a "guidance network" to enhance token decoding during inference. The pseudocode is detailed in Algorithm 1.

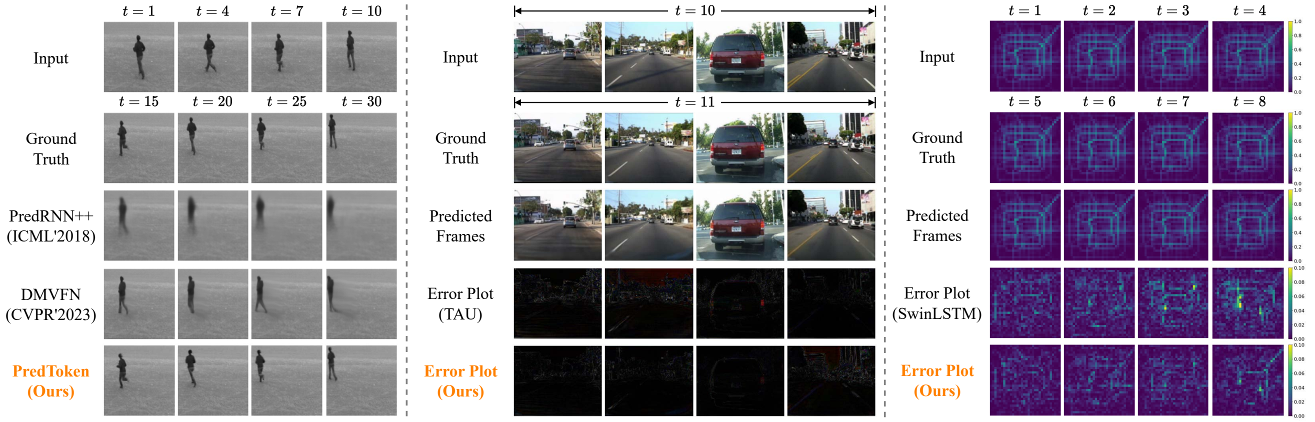
The overall inference process is shown in Figure 3.2 Stage III, dynamic soft decoding for coarse-grain token refinement, which then cascades into static soft decoding for fine-grain token refinement. In practice, the coarse level requires more iterations than the fine level.

Algorithm 1 Dynamic Soft Decoding

```
Input: input tokens $z$, mask $m$, steps $K$, temperature $T$
Output: output decoding tokens $\hat{z}$
$z_t^m \leftarrow \mathbf{mask}(z, m)$
for $t \leftarrow 0, 1, \ldots, K-1$ do
    $k \leftarrow \lceil\gamma\left(\frac{t+1}{K}\right)\cdot N\rceil$
    $\hat{z}_i \sim Soft(p_\theta(z_i \mid z_t^m))$
    $s_i \leftarrow p_\varepsilon(m_t^i \mid \hat{z}) + T(1 - \frac{t+1}{K})\,\mathrm{Gumbel}(0,1)$
        On non-mask indices of $z^m$: $s_i \leftarrow 1$
    $\tau \leftarrow$ The $k$-th smallest value of $s$
    $m_{t+1}^i \leftarrow \texttt{True}$ if $s_i < \tau$, $\texttt{False}$
    $z_{t+1}^m \leftarrow \mathbf{mask}(\hat{z}, m_{t+1})$
end for
return $\hat{z} = [\hat{z}_1, \hat{z}_2, \cdots, \hat{z}_N]$
```

4. Experiments

In this section, we qualitatively and quantitatively evaluate our method with prior works across diverse real-world benchmarks for predictive learning, including TaxiBJ [51],



(a) KTH (b) Kitti&Caltech (c) TaxiBJ

Figure 5. Qualitative results on the KTH (10 → 20), Kitti&Caltech (10 → 1) and TaxiBJ (4 → 4) datasets, where error plot = |ground truth − prediciton| denotes the differences between the ground truth frames and their corresponding predicted frames.

WeatherBench [25], KTH [27], Kitti&Caltech [7, 13], UCF101&DAVIS [23, 31], and SJTU4K [30]. The detailed statistics of benchmark datasets are in Table 1.

Implementation Details. We implement the proposed method with the PyTorch framework and trained on 4 NVIDIA A100 GPUs. The model is trained with a mini-batch of 16 video sequences, utilizing the Adam optimizer, a learning rate of 0.01, a weight decay of 0.05, a dropout rate of 0.1, and a shared codebook size of 16,384. The CGF and FGT are a stack of L_1 and L_2 blocks, where we use learnable spatiotemporal positional embeddings. The details of hyperparameters can be found in our appendix.

Table 1. The detailed statistics of each benchmark dataset.

Dataset	Size		Seq. Len.		Img. Shape	Interval
	train	test	in	out	H × W × C	
TaxiBJ [51]	20,461	500	4	4	32 × 32 × 2	30 min
WeatherBench [25]	2,167	706	12	12	32 × 64 × 1	1 hour
KTH [27]	4,940	3,030	10	20	128 × 128 × 1	Frame
Kitti&Caltech [7, 13]	3,160	3,095	10	1	128 × 160 × 3	Frame
UCF101&DAVIS [23, 31]	9,537	30	4	4	480 × 480 × 3	Frame
SJTU4K [30]	3,873	445	4	4	2160 × 3840 × 3	Frame

4.1. Comparison to State-of-the-Arts

In our evaluation, we benchmark our proposed PredToken against state-of-the-art spatiotemporal predictive models, including competitive recurrent-based architectures, *e.g.*, ConvLSTM [28], PredRNN [40], E3D-LSTM [42], PredRNNv2 [44], SwinLSTM [35] and recurrent-free architectures, *e.g.*, SimVP [10], TAU [33], DMVFN [18].

Predicting real-world traffic flow and forecasting weather are critical for public safety and scientific research. The TaxiBJ [51] and WeatherBench [25] evaluate these models on a macro scale, yet they exhibit lower frequencies compared to other tasks. Therefore, predictive models must detect subtle changes, with quantitative outcomes presented in Table 2. Qualitative visualizations can be found in Figure 5(c) and Figure 7. Notably, our PredToken method outperforms others by generating the sparsest error plot, demonstrating superior capability in capturing traffic and climate patterns.

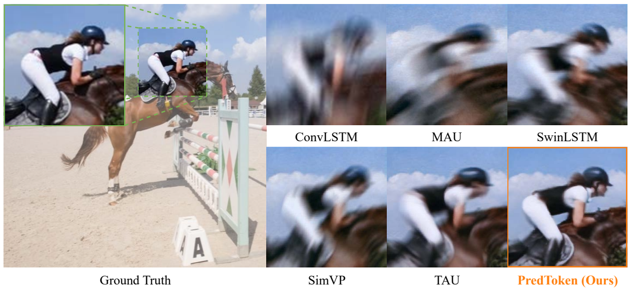


Figure 6. The visualization of (t+4)-th frame results between our PredToken and the state-of-the-art methods on the DAVIS17-Val.

Table 2. Quantitative results on the TaxiBJ (4 → 4 frames) and WeatherBench (12 → 12 frames) datasets, higher SSIM and PSNR, lower MAE and RMSE indicate better results.

Method	TaxiBJ		WeatherBench	
	SSIM↑	PSNR↑	MAE↓	RMSE↓
ConvLSTM (NIPS'2015) [28]	0.978	37.38	0.7949	1.233
PredRNN++ (ICML'2018) [41]	0.977	38.71	0.7883	1.278
MIM (CVPR'2019) [43]	0.971	38.71	0.8716	1.336
E3D-LSTM (ICLR'2019) [42]	0.979	38.75	0.8059	1.233
SimVP (CVPR'2022) [10]	0.982	39.17	0.7037	1.113
PredRNNv2 (PAMI'2022) [44]	0.983	39.38	0.7986	1.243
TAU (CVPR'2023) [33]	0.982	39.50	0.6810	1.106
SwinLSTM (ICCV'2023) [35]	0.977	38.71	0.7220	1.130
Ours	**0.985**	**39.79**	**0.6347**	**1.042**

Predicting human motion and driving dynamics is challenging due to the significant variability in individual behaviors and actions. Following the standard settings [33],

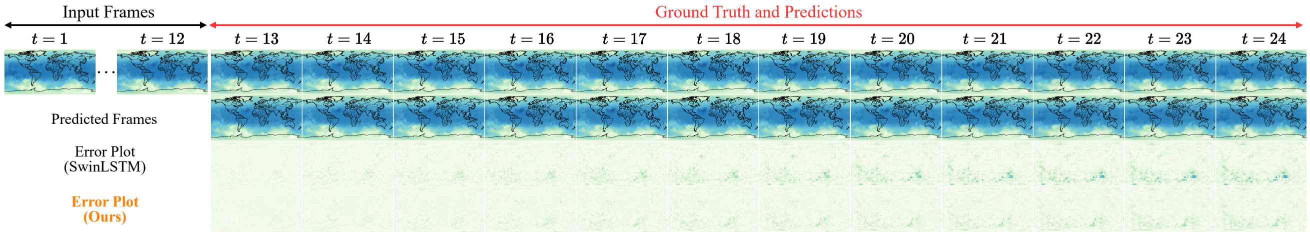


Figure 7. Qualitative results on the WeatherBench (12 → 12) for global temperature forecasting at 5.625° resolution (32×64 grid points).

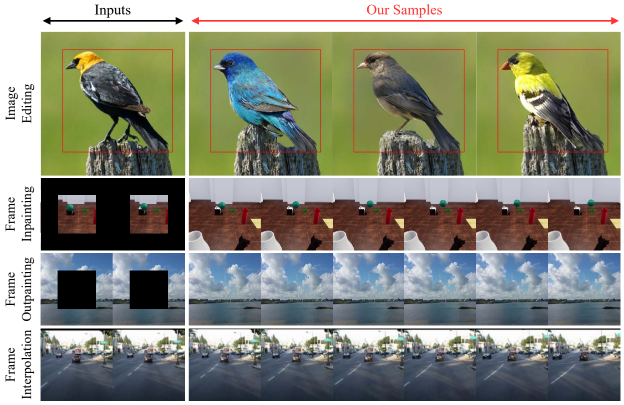


Figure 8. The visualization of more visual generative tasks on ImageNet [6], Physion [1], Sky Time-lapse [50], and KITTI [13].

we employ the KTH [27] and Kitti&Caltech [7, 13] to evaluate the performance of the models in these scenarios. The quantitative results are reported in Table 3. PredToken achieves state-of-the-art results, notably in LPIPS with a reduction from 0.2394 to 0.1265 compared with TAU [33] on the KTH. In Figure 5(a), PredRNN++ [41] predicts accurate position but blurs around the human body, and DMVFN [18] produces a sharp body but deviates in position and action from the ground truth. Our PredToken excels in precise position predictions, sharp visual representations, and faithful actions. Furthermore, as illustrated in Figure 5(b), our method produces clearer predictions in vehicle dynamics and lane lines compared to other methods.

Table 3. Quantitative results on the KTH (10 → 20 frames) and Kitti&Caltech (10 → 1 frames) datasets, higher PSNR, lower MSE and LPIPS indicate better results.

Method	KTH		Kitti&Caltech	
	PSNR↑	LPIPS↓	MSE↓	PSNR↑
ConvLSTM (NIPS'2015) [28]	23.58	0.5128	139.6	27.46
PredRNN (NIPS'2017) [40]	27.55	0.4621	130.4	27.81
PredRNN++ (ICML'2018) [41]	28.47	0.4724	129.6	27.89
E3D-LSTM (ICLR'2019) [42]	29.31	0.4835	200.6	25.45
SimVP (CVPR'2022) [10]	33.72	0.2649	160.2	26.81
DMVFN (CVPR'2023) [18]	32.15	0.1284	183.9	26.78
TAU (CVPR'2023) [33]	34.13	0.2394	131.1	27.83
Ours	**35.11**	**0.1265**	**118.9**	**28.81**

Table 4. Quantitative results on the UCF101&DAVIS (4 → 4 frames) and SJTU4K (4 → 4 frames) datasets.

Method	UCF101&DAVIS				SJTU4K			
	SSIM $\times 10^{-2}$ ↑		LPIPS $\times 10^{-2}$ ↓		PSNR ↑		LPIPS $\times 10^{-2}$ ↓	
	t+1	t+3	t+1	t+3	t+1	t+3	t+1	t+3
ConvLSTM [28]	68.81	55.97	23.42	34.51	22.74	17.91	67.81	86.84
PredRNN [40]	78.78	70.26	13.12	21.56	23.25	18.20	66.60	87.04
SimVP [10]	83.96	74.35	10.31	17.21	24.57	20.17	56.42	66.03
TAU [33]	84.81	75.05	9.41	16.24	25.68	21.03	55.84	65.21
SwinLSTM [35]	83.76	74.13	10.44	17.78	24.37	19.77	57.12	66.68
Ours	**88.47**	**78.34**	**7.52**	**13.38**	**28.71**	**23.89**	**51.41**	**61.37**

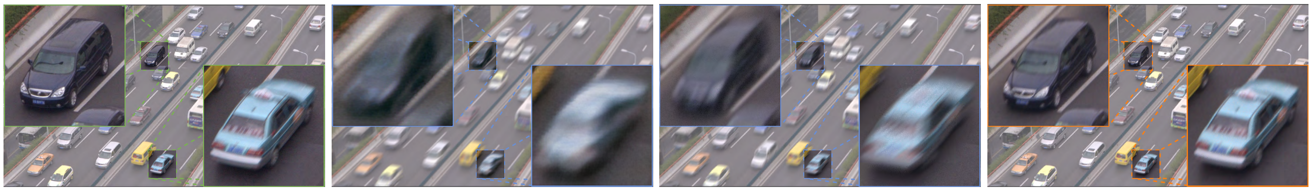
High-resolution prediction in the real world is notably challenging due to the complexity of textures and motions. Following the established settings [5, 18], we utilized the UCF101&DAVIS and SJTU4K datasets. For UCF101&DAVIS, PredToken, trained on UCF101 [31] and evaluated on DAVIS17-Val [23], surpasses other methods, setting new benchmarks as indicated in the Table 4. For the SJTU4K, PredToken significantly improves PSNR (21.03 → 23.89) and reduces LPIPS (65.21 → 61.37) at (t+3)-th frame compared with TAU [33]. We provide qualitative visualization examples in Figure 6 and Figure 9. PredToken excels in capturing the intricate dynamics of scenes, especially in predicting human and vehicle motions, offering sharp visuals and strong generalization.

4.2. More Visual Generative Tasks

To explore the versatility of PredToken in diverse visual generative tasks, we conducted class-conditional image editing, frame inpainting, frame outpainting, and frame interpolation tasks, as illustrated in Figure 8. For image editing, we treat an image as a single-frame video for PredToken, conducting category-based edits via specific masked areas. For frame inpainting and outpainting, we apply masks outside and inside the frame, respectively, and for frame interpolation, masks are positioned between frames. The experimental outcomes show that PredToken can adapt to diverse generative tasks, producing realistic results.

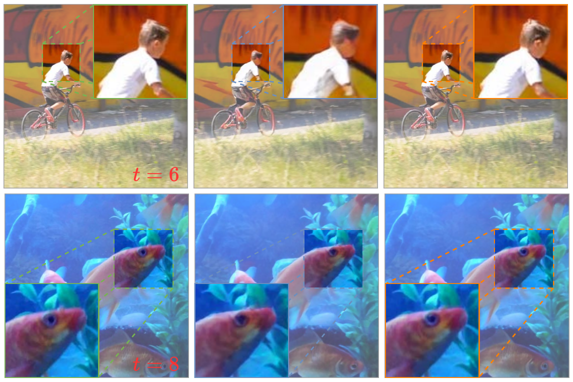
4.3. Ablation Study

Tokenizer Reconstruction Quality. We indirectly evaluate the learned token representation through tokenizer reconstruction experiments. Table 5 shows quantitative re-



Ground Truth SwinLSTM (ICCV'2023) TAU (CVPR'2023) **PredToken (Ours)**

Figure 9. The visualization of (t+1)-th frame results on the SJTU4K dataset (4 → 4 frames) at a resolution of 2160 × 3840 pixels.



Ground Truth w/o DQR (VQGAN) **w/ DQR**

Figure 10. The visualization of tokenizer reconstruction quality for 8 frames on DAVIS17 dataset (480 × 480 resolution).

Table 5. Ablation studies of tokenizer reconstruction quality on DAVIS17 and iterative decoding method on KTH.

Reconstruction Quality					Iterative Decoding			
DQR	DQM	SC	SSIM↑	PSNR↑	SD	MD	SSIM↑	PSNR↑
✗	✗	✗	80.4	28.1	✗	✗	87.8	31.97
✓	✗	✗	89.3 (+8.9)	30.9 (+2.8)	✓	✗	88.5 (+0.7)	32.48 (+0.51)
✓	✓	✗	89.7 (+9.3)	31.5 (+3.4)	✗	✓	89.4 (+1.6)	32.94 (+0.97)
✓	✓	✓	89.8 (+9.4)	31.7 (+3.6)	✓	✓	89.8 (+2.0)	33.03 (+1.06)

sults. Compared to the baseline VQGAN, it is evident that using our proposed "decomposition, quantization, and reconstruction" (DQR) framework results in significant improvements in reconstruction quality, which highlights the effectiveness of learning from disentangled low- and high-frequency representations. We also observe performance gains by adding a dimension-aware quantization module (DQM) ahead of the codebook. The shared codebook (SC) also contributes to model performance. Furthermore, We visualized 8-frame reconstruction examples on the DAVIS17 dataset, as shown in Figure 10, noting the vanilla VQGAN without DQR (w/o DQR) misses key details, *e.g.*, head details and textures or fish eyes.

Iterative Decoding Method. We evaluate the proposed iterative decoding methods on the KTH. The models predict 40 future frames from 10 past observations. Table 5 details their impact during the coarse-level token generation stage. Compared to the original iterative decoding, incorporating soft decoding (SD) can enhance performance during inference. The mask discriminator (MD), unlike previous discriminators, dynamically guides the token decoding process, resulting in notable gains. We also explored the order of static soft decoding (SSD) and dynamic soft decoding (DSD), as shown in Figure 11. The results indicate the combination of DSD for the coarse level and SSD for the fine level is most effective. While using DSD + DSD is theoretically optimal, its extra training objectives slow down the overall loss convergence. We also presented comparisons with other methods for long-term prediction tasks, highlighting that while the performance of other methods declines rapidly with increasing frame numbers, our approach exhibits a slower decline. This demonstrates the superior long-term predictive capability of PredToken.

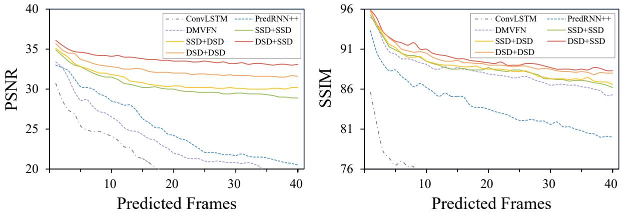


Figure 11. Ablation studies of decoding methods combination and quantitative comparison with prior work [18, 28, 41] on KTH.

5. Conclusion

This paper presents PredToken, an innovative spatiotemporal predictive framework that highlights low-level consistency and high-level dynamics. First, we introduce a "decomposition, quantization, and reconstruction" schema based on VQGAN to enhance token representation, effectively separating frequency components and utilizing a dimension-aware quantization model to preserve intricate details. Second, we propose a "coarse-to-fine iterative decoding" method to improve token decoding, which leverages dynamic and static soft decoding for coarse and fine tokens, respectively, enabling more high-level dynamics to be captured. Extensive experiments show the superiority of our method in various real-world predictive benchmarks, and PredToken can be adapted for other visual generative tasks to yield realistic results.

References

[1] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. 7

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[3] 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. 2

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 3, 5

[5] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Strpm: A spatiotemporal residual predictive model for high-resolution video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13946–13955, 2022. 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, pages 304–311. IEEE, 2009. 6, 7

[8] 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. 2, 3, 4

[9] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022. 3

[10] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3170–3180, 2022. 3, 6, 7

[11] Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning, 2024. 1

[12] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 2

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6, 7

[14] Ian 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. 5

[15] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022. 2, 3

[16] Weizhen He, Weijie Chen, Binbin Chen, Shicai Yang, Di Xie, Luojun Lin, Donglian Qi, and Yueting Zhuang. Unsupervised prompt tuning for text-driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2661, 2023. 1

[17] 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. 1

[18] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2023. 3, 6, 7, 8

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[20] 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. 3

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2

[22] Xuesong Nie, Xi Chen, Haoyuan Jin, Zhihang Zhu, Yunfeng Yan, and Donglian Qi. Triplet attention transformer for spatiotemporal predictive learning. *arXiv preprint arXiv:2310.18698*, 2023. 1, 2

[23] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6, 7

[24] Shengju Qian, Huiwen Chang, Yuanzhen Li, Zizhao Zhang, Jiaya Jia, and Han Zhang. Strait: Non-autoregressive generation with stratified image transformer. *arXiv preprint arXiv:2303.00750*, 2023. 3

[25] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12 (11):e2020MS002203, 2020. 6

[26] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 4

[27] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, pages 32–36. IEEE, 2004. 6, 7

[28] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 2, 6, 7, 8

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[30] Li Song, Xun Tang, Wei Zhang, Xiaokang Yang, and Pingjian Xia. The sjtu 4k video sequence dataset. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 34–35. IEEE, 2013. 6

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012. 6, 7

[32] Jiahao Su, Wonmin Byeon, Jean Kossaifi, Furong Huang, Jan Kautz, and Anima Anandkumar. Convolutional tensor-train lstm for spatio-temporal learning. *Advances in Neural Information Processing Systems*, 33:13714–13726, 2020. 2

[33] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18782, 2023. 1, 3, 6, 7

[34] Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z Li. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. *arXiv preprint arXiv:2306.11249*, 2023. 1

[35] Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. *arXiv preprint arXiv:2308.09891*, 2023. 2, 6, 7

[36] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 1, 3

[37] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1, 3

[38] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021. 1, 3

[39] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024. 2, 3

[40] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017. 2, 6, 7

[41] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and S Yu Philip. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International Conference on Machine Learning*, pages 5123–5132. PMLR, 2018. 6, 7, 8

[42] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *International conference on learning representations*, 2018. 2, 6, 7

[43] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9154–9162, 2019. 6

[44] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, S Yu Philip, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022. 6

[45] Ziru Xu, Yunbo Wang, Mingsheng Long, Jianmin Wang, and M KLiss. Predcnn: Predictive learning with cascade convolutions. In *IJCAI*, pages 2940–2947, 2018. 2

[46] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

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

[48] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 2

[49] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 2, 3, 5

[50] Jiangning Zhang, Chao Xu, Liang Liu, Mengmeng Wang, Xia Wu, Yong Liu, and Yunliang Jiang. Dtvnet: Dynamic time-lapse video generation via single still image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 300–315. Springer, 2020. 7

[51] Junbo Zhang et al. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 5, 6

[52] Song Zhang, Qingzhong Wang, Jie Fu, Jiang Bian, and Haoyi Xiong. Vision electra: Adversarial masked image modeling with hierarchical discriminator. 2023. 3

[53] Yiqi Zhong, Luming Liang, Ilya Zharkov, and Ulrich Neumann. Mmvp: Motion-matrix-based video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4273–4283, 2023. 2