A Survey on Future Frame Synthesis: Bridging Deterministic and Generative Approaches

Anonymous authors
Paper under double-blind review

Abstract

Future Frame Synthesis (FFS) focuses on generating future frame sequences conditioned on existing content. This survey provides a comprehensive review of existing research on FFS, covering commonly used datasets and representative algorithms. We discuss key challenges and trace the evolution of FFS in computer vision, particularly the shift from deterministic to generative approaches. Our taxonomy outlines major advances and methodological shifts, emphasizing the rising significance of generative models in producing realistic and diverse predictions.

1 Introduction

The goal of the Future Frame Synthesis (FFS) task is to generate future frames from a sequence of historical frames (Srivastava et al., 2015) or even a single context frame (Xue et al., 2016), optionally guided by supplementary control signals. The learning objective of FFS is also regarded as central to building a world model (Ha & Schmidhuber, 2018a; Hafner et al., 2023). FFS is closely related to low-level computer vision techniques, particularly when synthesizing temporally adjacent frames (Liu et al., 2017; Wu et al., 2022b; Hu et al., 2023b). However, FFS differs from other low-level tasks by implicitly requiring a more sophisticated understanding of scene dynamics and temporal coherence—traits typically associated with high-level vision tasks. The key challenge is to design models that can achieve this balance efficiently, using a moderate number of parameters to reduce inference latency and resource consumption, thereby making FFS practical for real-world deployment. This unique positioning underscores the integral role of FFS in bridging the gap between low-level perception & prediction, and high-level understanding & generation in computer vision.

Contemporary FFS algorithms are generally categorized into two main approaches. The first approach involves **referencing** pixels from existing frames—typically the last observed frame—to synthesize future content. However, such methods inherently struggle to model the appearance and disappearance of objects. These methods tend to produce accurate short-term predictions but degrade over longer time horizons. This line of research is commonly known as video prediction (Oprea et al., 2020). The second approach focuses on **generating** future frames from scratch. While these methods hold the potential to model object emergence and disappearance, they primarily operate at the pixel level. As a result, they often fail to capture high-level semantic context, which is essential for realistic and imaginative generation.

Prior to our work, two surveys published in 2020 (Oprea et al., 2020; Rasouli, 2020) provided comprehensive overviews of early technical developments in video prediction. More recently, several surveys have emerged on text-to-video generative models (Liu et al., 2024) and long video generation (Li et al., 2024; Sun et al., 2024b). In contrast, our survey emphasizes recent advances and explores the interplay between predictive and generative methodologies. We argue that the future of long-term FFS lies in the synergistic integration of prediction and generation techniques. Such a unified approach integrates contextual constraints with semantic understanding, enabling more robust and coherent synthesis.

As a foundation, we introduce the problem formulation and core challenges in Section 2. Our taxonomy is organized around the degree of stochasticity in modeling approaches. In Section 3, we present deterministic algorithms that aim to perform pixel-level fitting based on fixed target frames. However, pixel-level metrics tend to drive models to average over multiple plausible futures, often resulting in blurry predictions. In

Section 4, we examine algorithms that enable stochastic motion prediction. These include approaches that inject stochastic variables into deterministic models, as well as methods based on explicit probabilistic modeling. Such methods enable sampling from motion distributions, yielding diverse yet plausible predictions beyond the target frame. Given the limited generative capacity of current FFS algorithms—especially for high-resolution videos involving object appearance and disappearance—we introduce the generative FFS task in Section 5. This task prioritizes producing coherent long-term sequences over pixel-level accuracy. In Section 6, we explore the broad applicability of FFS in areas such as world model, autonomous driving, robotics, film production, meteorology, and anomaly detection. These use cases demonstrate the role of FFS in dynamic scene understanding and interaction. In Section 7, we review prior surveys on video prediction and diffusion-based video generation. We also clarify our distinct focus: a comprehensive analysis of FFS spanning from deterministic to generative paradigms, emphasizing the growing role of generative models in producing realistic and diverse future predictions.

2 Future frame synthesis

2.1 Problem Definition

The FFS task involves predicting future frames conditioned on previously observed video content. The primary objective is to develop models capable of accurately capturing future visual dynamics. Formally, this task can be formulated as a conditional generative modeling problem: given observed frames $X_{t_1:t_2}$, the goal is to generate future frames $Y_{t_2+1:t_3}$. This relationship can be expressed as the conditional probability distribution:

$$Y_{t_2+1:t_3} \sim \mathbb{P}(Y_{t_2+1:t_3} \mid X_{t_1:t_2}), \tag{1}$$

In Eq. (1), t_1 denotes the initial time step, t_2 marks the end of the observed frame sequence, and t_3 indicates the final time step for future frame synthesis. The key challenge is to learn a mapping function that models the complex spatio-temporal dependencies across frames. Here, $\mathbb{P}(Y_{t_2+1:t_3} \mid X_{t_1:t_2})$ denotes the conditional probability distribution over future frames given the observed sequence.

Many FFS algorithms incorporate additional information, such as auxiliary data $A_{t_1:t_2}$ extracted from videos—including depth maps, landmarks, bounding boxes, and segmentation maps—to enhance scene understanding. They may also include human control signals $C_{t_2+1:t_3}$, such as text instructions or sketch-based strokes, which guide the model to generate future sequences following specific intended trajectories. Taking these factors into account, we extend the FFS formulation to a more comprehensive version, as shown in Eq. (2):

$$Y_{t_2+1:t_3} \sim \mathbb{P}(Y_{t_2+1:t_3} \mid X_{t_1:t_2}, A_{t_1:t_2}, C_{t_2+1:t_3}) \tag{2}$$

2.2 Method Paradigms

In the domain of Future Frame Synthesis (FFS), three paradigms have emerged—deterministic, stochastic, and generative—each representing a distinct modeling approach. The deterministic paradigm emphasizes pixel-level fitting to fixed target frames, typically employing low-level computer vision architectures such as Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012) and Recurrent Neural Networks (RNNs) (Rumelhart et al., 1985). However, optimizing pixel-level metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) (Wang et al., 2004) often leads to blurry outputs, as the models tend to average over multiple plausible futures. In contrast, the stochastic paradigm introduces randomness into the prediction process by incorporating stochastic variables or distributions to model the inherent uncertainty in video dynamics. Another approach involves using probabilistic models, such as Variational Autoencoders (VAEs) (Kingma & Welling, 2014) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), which can produce diverse future frames that deviate from the ground truth while remaining plausible. The generative paradigm, by contrast, prioritizes the synthesis of coherent and plausible video sequences over pixel-level fidelity. It leverages advanced generative models, such as diffusion models and large language models, to produce diverse and imaginative future frames that capture complex scene dynamics, including object emergence and disappearance. As research evolves, the boundaries between these paradigms continue to blur, with growing efforts to integrate their strengths and build more capable and versatile FFS systems.

2.3 Overall Challenges

The field of FFS faces several longstanding challenges, including the need for algorithms that balance low-level pixel fidelity with high-level scene understanding, the lack of reliable perceptual and stochastic evaluation metrics, the difficulty of achieving long-term synthesis, and the scarcity of high-quality high-resolution datasets that capture stochastic motion and object emergence and disappearance. This section outlines these key challenges and sets the stage for further discussion.

2.3.1 Evaluation metrics

Low-level metrics such as PSNR and SSIM assess only the pixel-wise accuracy of predictions. To optimize for these metrics, models are typically trained using pixel-space losses such as ℓ_1 or ℓ_2. This often leads to blurry predictions that closely match the ground truth, rather than sharper and more realistic generations that deviate from it—a phenomenon known as the perception-distortion trade-off (Blau & Michaeli, 2018). As a result, researchers are increasingly exploring alternative evaluation metrics, including perceptual metrics (e.g., DeePSiM (Dosovitskiy & Brox, 2016), LPIPS (Zhang et al., 2018)) and stochastic metrics (e.g., IS (Salimans et al., 2016), FID (Heusel et al., 2017)). These metrics are believed to better align with human perceptual judgments. However, even classifiers trained on human-annotated perceptual data show limited agreement with human judgments of image quality (Kumar et al., 2022).

In visual domains, models are typically evaluated based on the perceptual quality of their generated outputs. However, for many practical applications, perceptual quality may not be the most critical factor. For instance, Dreamer-V3 (Hafner et al., 2023) and VPT (Baker et al., 2022) have successfully built effective world models using low-resolution frame sequences. Moreover, most visual representation learning methods are developed using relatively low-resolution images (Radford et al., 2021; He et al., 2022). We are concerned that an excessive pursuit of visual quality may bias model selection toward architectures that overfit low-level features. Beyond aligning with human perceptual judgments, evaluation metrics should also be designed to assess a model's capacity to capture scene dynamics and temporal variations.

Even with improved evaluation metrics, optimizing them during training remains a significant challenge. During training, researchers often use pretrained ImageNet classifiers as feature extractors (Johnson et al., 2016; Kumar et al., 2022) to compare generated outputs with ground truth, thereby optimizing for both low-level and high-level features. Additionally, GAN-based loss functions have been proposed to enhance the perceptual quality of generated outputs (Huang et al., 2017; Zhang et al., 2020a).

2.3.2 Long-term synthesis

Despite significant progress in short-term video prediction, synthesizing events over extended time horizons remains challenging due to long-term dependencies and complex object interactions in dynamic scenes. Naively applying short-term models in an iterative manner often leads to rapid quality degradation (Wu et al., 2022b; Hu et al., 2023b). Owing to limited model capacity for forming a comprehensive understanding of the real world, most existing video synthesis models primarily model pixel-level distributions. When dealing with natural videos over long durations, these models struggle to predict object dynamics while preserving visual quality. One promising direction is to incorporate high-level structural information (Villegas et al., 2017). Leveraging such higher-order representations helps models retain key details and maintain temporal consistency over extended time scales.

2.3.3 Generalization

The interplay between data volume and model complexity jointly determines the upper bounds of an algorithm's performance. Despite the vastness of video data available on the internet, the scarcity of high-quality

Dataset	Category	# Videos	# Clip Frames	Resolution	Extra Annotations
KTH Action (Schuldt et al., 2004)	Human	2, 391	95*	160 × 120	Class
Caltech Pedestrian (Dollar et al., 2011)	Human	137	1, 824	640 × 480	Bounding Box
HMDB51 (Kuehne et al., 2011)	Human	6, 766	93*	414 × 404†	Class
UCF101 (Soomro et al., 2012)	Human	13, 320	187*	320 × 240	Class
J-HMDB (Jhuang et al., 2013)	Human	928	34*	320 × 240	OF, Ins, HJ, Class
KITTI (Geiger et al., 2013)	Traffic	151	323*	1242 × 375	OF, BBox, Sem, Ins, Depth
Penn Action (Zhang et al., 2013)	Human	2, 326	70*	480 × 270†	Human Joint, Class
SJTU 4K (Song et al., 2013)	General	15	300	3840 × 2160	-
Sports-1M (Karpathy et al., 2014)	Human	1, 133, 158	variable	variable	Class
Moving MNIST (Srivastava et al., 2015)	Simulation	10, 000	20	64 × 64	-
Cityscapes (Cordts et al., 2016)	Traffic	46	869*	2048 × 1024	Semantic, Instance, Depth
YouTube-8M (Abu-El-Haija et al., 2016)	General	8, 200, 000	variable	variable	Class
Robotic Pushing (Finn et al., 2016)	Robot	59, 000	25*	640 × 512	Class
DAVIS17 (Pont-Tuset et al., 2017)	General	150	73*	3840 × 2026†	Semantic
Something-Something (Goyal et al., 2017)	Object	220, 847	45	427 × 240†	Text
ShapeStacks (Groth et al., 2018)	Simulation	36, 000	16	224 × 224	Semantic
SM-MNIST (Denton & Fergus, 2018)	Simulation	customize	customize	64 × 64	-
D^2-City (Che et al., 2019)	Traffic	11, 211	750*	1080p / 720p	BBox
Kinetics-700 (Carreira et al., 2019)	Human	650, 000	250*	variable	Class
RoboNet (Dasari et al., 2019)	Robot	161, 000	93*	64 × 48	-
Vimeo-90K (Xue et al., 2019)	General	91, 701	7	448 × 256	-
BDD100K (Yu et al., 2020)	Traffic	100, 000	1175*	1280 × 720	BBox, Semantic, Depth
nuScenes (Caesar et al., 2020)	Traffic	1, 000	40*	1600 × 900	BBox, Semantic
WebVid (Bain et al., 2021)	General	10, 732, 607	449*	596 × 336	Text
X4K1000FPS (Sim et al., 2021)	General	4, 408	65*	4096 × 2160	-
SportsSlomo (Chen & Jiang, 2024)	Human	130,000	7	1280 × 720	-
InternVideo2 (Wang et al., 2024d)	General	2, 000, 000	variable	variable	Text, Action
HowTo100M (Miech et al., 2019)	General	136, 600, 000	variable	variable	Text
OpenDV-YouTube (Yang et al., 2024)	Traffic	2, 139	100,000+*	variable	Text

* denotes the mean value. † denotes the median value.

Table 1: Summary of commonly used video prediction datasets, including dataset category, total number of videos, frame count for each video clip, image resolution, and additional annotations. (**OF**: Optical Flow, **BBox**: Bounding Box, **Sem**: Semantic, **Ins**: Instance, **HJ**: Human Joints)

video datasets suitable for video synthesis remains a limiting factor.Existing datasets often suffer from simplistic data distribution, low resolution, and limited motion diversity. These limitations hinder the ability of current video synthesis models to handle high-resolution content and large motion scales, thus restricting their practical utility to diverse and unseen scenarios. High-resolution video synthesis is inherently challenging and requires substantial computational resources (Blattmann et al., 2023b). The computational burden further complicates real-time deployment in practical applications.

2.4 Datasets

The advancement of video synthesis models heavily depends on the diversity, quality, and characteristics of training datasets. A common observation is that the suitability of datasets varies with their dimensionality and size: lower-dimensional datasets, often smaller in scale, tend to exhibit limited generalization. In contrast, higher-dimensional datasets offer greater variability, contributing to stronger generalization capabilities. In Table 1, we summarize the most widely used datasets in video synthesis, highlighting their scale and available supervisory signals to provide a comprehensive overview of the current dataset landscape. For datasets lacking detailed reports in their original papers or project pages, we estimate missing values using mean or median statistics to ensure consistency across the analysis.

Challenges. 1. Unifying the organization of image and video data. A large proportion of computer vision research has historically focused on the image modality. As a result, image datasets are often more carefully curated and contain richer annotations. Representative large-scale image datasets include YFCC100M (Thomee et al., 2016), WIT400M (Radford et al., 2021), and LAION400M (Schuhmann et al., 2021). Given the scale of available image data, it is important to effectively leverage knowledge from foundational image models. When incorporating video data into training pipelines, it is often necessary to filter out low-quality segments and select an appropriate sampling frame rate.

2. Determining the proportion of data from different domains. Computer graphics composite data, 2D anime data, real-world videos, and videos with special effects exhibit vastly different visual characteristics. Moreover, standardizing data from diverse sources to a fixed resolution is challenging due to varying aspect ratios and the presence of resolution-dependent details such as subtitles and textures. Many frame synthesis methods are sensitive to resolution partly due to the correlation between resolution and object motion intensity (Sim et al., 2021; Hu et al., 2023b; Yoon et al., 2024).

3 Deterministic Synthesis

Method	Publication	Main Ideas
ConvLSTM (Shi et al., 2015)	NeurIPS'15	Formulate precipitation nowcasting as a spatio-temporal sequence forecasting problem, propose the convolutional LSTM to build an end-to-end model.
PredNet (Lotter et al., 2017)	ICLR'17	Use a recurrent CNN with both bottom-up and top-down connections, with each neural layer making local residual predictions.
PredRNN (Wang et al., 2017)	NeurIPS'17	Introduce a new spatio-temporal LSTM unit, that extracts and memorizes spatial and temporal representations simultaneously.
E3d-LSTM (Wang et al., 2019)	ICLR'19	Integrate 3D convolutions into RNNs to enhance local motion modeling, and make the present memory state interact with its long-term historical records.
MSPred (Villar-Corrales et al., 2022)	BMVC'22	Focus on enhancing long-term action planning ability, and use spatiotemporal downsampling to forecast at different scales.
Multi-Scale AdvGDL (Mathieu et al., 2016)	ICLR'16	Propose three feature learning strategies to address blurriness in FFS: a multi-scale architecture, adversarial training, and an image gradient difference loss.
PredCNN (Xu et al., 2018)	IJCAI'18	Design a cascade multiplicative unit that provides more operations for previous frames, and capture the temporal dependencies through stacked operations.
DVF (Liu et al., 2017)	ICCV'17	Learn to synthesize video frames by warping pixel values from existing ones, and unify video interpolation and extrapolation within a single CNN framework.
SimVP (Gao et al., 2022)	CVPR'22	Propose a simple CNN video prediction model without complicated tricks, and provide insights for selecting different architectures.
SDC-Net (Reda et al., 2018)	ECCV'18	Learn a motion vector and a kernel for each pixel to synthesize the predicted pixel, and inherit the merits of both vector-based and kernel-based approaches.
FVS (Wu et al., 2020)	CVPR'20	Decouple the background scene and moving objects, construct future frames using non-rigid deformation of the background and affine transformation of objects.
OPT (Wu et al., 2022b)	CVPR'22	Solve FFS as an optimization problem, with a pretrained VFI module to construct function. By eliminating the domain gap, OPT is robust in general scenarios.
DMVFN (Hu et al., 2023b)	CVPR'23	On the basis of DVF, multi-scale coarse-to-fine prediction is added, and the input is processed by a dynamic routing subnetwork at the inference stage.
VPTR Ye & Bilodeau (2023)	IVC'23	Present a new efficient building block of transformer-based models for FFS, along with three competitive variants for video prediction.
S2S (Luc et al., 2017)	ICCV'17	Introduce the task of predicting semantic segmentations of future frames, and show that predicting future segmentations is substantially better than segmenting predicted frames.
SADM (Bei et al., 2021)	CVPR'21	Decompose the scene layout (semantic map) and motion (optical flow) into layers, and hope to explicitly represent objects and learn their class-specific motion.
MAL (Liu et al., 2023)	WACV'23	Predict the depth maps of future frames using a two-branch structure. One branch handles future depth estimation, and the other aids image reconstruction.

Table 2: Overview of deterministic synthesis methods.

3.1 Raw Pixel Space

In short-term FFS, approaches operating in the raw pixel space have achieved promising results. In this section, we review representative methods and discuss the associated challenges.

3.1.1 Recurrent networks.

PredNet (Lotter et al., 2017) pioneered the exploration of recurrent neural networks in video synthesis, drawing inspiration from predictive coding in neuroscience and employing a recurrent convolutional network to process video features effectively. Building on this foundation, PredRNN (Wang et al., 2017) introduces significant improvements by modifying the Long Short-Term Memory (LSTM) architecture with a dual-memory structure, aiming to enhance spatio-temporal modeling. Despite these advances, the model still faces challenges such as gradient vanishing in video synthesis tasks. To address these limitations, ConvLSTM (Shi et al., 2015) emerges as a pivotal model by ingeniously integrating LSTM with CNNs to effectively capture motion and spatio-temporal dynamics—a development that has significantly influenced subsequent video synthesis models. E3d-LSTM (Wang et al., 2019) further advances the field by incorporating 3D convolutions into RNNs and introducing a gate-controlled self-attention module, thereby significantly improving long-term synthesis capabilities. However, the increased computational complexity introduced by 3D convolutions may offset the performance gains in certain applications. MSPred (Villar-Corrales et al., 2022) proposes a hierarchical convolutional-recurrent network that operates at multiple temporal frequencies to predict future video frames, as well as other representations such as poses and semantics.

Challenges. Despite their effectiveness in capturing temporal dependencies, recurrent networks face several challenges in video prediction tasks. Their inherently sequential nature, which enables frame-by-frame modeling, can result in high computational complexity—particularly in high-resolution scenarios. This is evident from the significantly higher FLOPs and lower FPS observed in recurrent-based models compared to their recurrent-free counterparts (Tan et al., 2023). Additionally, recurrent networks are prone to gradient vanishing and exploding problems, which can severely hinder their ability to learn long-term dependencies (Gao et al., 2022). These challenges underscore the need for alternative approaches that balance efficiency and performance—such as recurrent-free models, which have demonstrated promising results across various video prediction tasks.

3.1.2 Convolutional networks.

CNNs play an instrumental role in the evolution of video synthesis technology. The progress began with Multi-Scale AdvGDL (Mathieu et al., 2016), initiating a series of significant advances in the field. Following this, PredCNN (Xu et al., 2018) establishes a new benchmark by outperforming its predecessor PredRNN (Wang et al., 2017) across various datasets. The introduction of SimVP (Gao et al., 2022) marks another milestone in convolutional approaches to video prediction. Inspired by the advances of Vision Transformers (ViT) (Dosovitskiy et al., 2021), this approach introduces a simplified CNN architecture to extract continuous tokens, demonstrating that such a configuration can achieve comparable performance in video synthesis.

Challenges. Although simple to implement and fast, CNN-based frame synthesis methods are not well-suited for spatially shifting the pixels of input frames. CAIN (Choi et al., 2020) and FLAVR (Kalluri et al., 2023) respectively introduce channel attention and 3D U-Net architectures for intermediate frame synthesis, but they do not fully replace explicit pixel motion approaches such as kernel-based and flow-based methods. Moreover, in pursuit of efficiency, most CNN-based models used for FFS maintain a relatively small parameter count, typically under 60M (Tan et al., 2023). By contrast, video diffusion models have been scaled up to over 1.5B parameters (Blattmann et al., 2023a) in order to fully leverage large-scale datasets. Effectively scaling up CNN-based models remains a significant challenge. It is speculated that short-term prediction models for high-resolution, real-time applications and those aiming to leverage large datasets for enhanced generation capabilities will follow diverging development paths.

3.1.3 Optical-flow-based synthesis.

Optical flow describes the motion of pixels between frames and can be used to warp pixels from the current frame to synthesize near-future frames (Liu et al., 2017). Flow-based methods can be viewed as extensions of kernel-based approaches (Niklaus et al., 2017), as the latter typically constrain pixel motion to a relatively small neighborhood (Cheng & Chen, 2021). SDC-Net (Reda et al., 2018) proposes a hybrid approach that inherits the strengths of both vector-based and kernel-based methods. FVS (Wu et al., 2020) enhances synthesis quality by incorporating supplementary information, such as semantic maps and instance maps from input frame sequences. While effective, this method introduces challenges due to increased data modalities and higher computational demands. OPT (Wu et al., 2022b) estimates optical flow through an optimization-based approach. By iteratively refining the current optical flow estimation, the quality of the next frame can be significantly improved. This approach effectively leverages knowledge from off-the-shelf optical flow models (Teed & Deng, 2020) and video frame interpolation methods (Jiang et al., 2018; Huang et al., 2022b).

Although training is not required, the iterative optimization process during inference incurs substantial computational cost. DMVFN (Hu et al., 2023b) improves dense voxel flow (Liu et al., 2017) estimation by dynamically adapting the network architecture based on motion magnitude. DMVFN further confirms the effectiveness of a coarse-to-fine, multi-scale strategy in addressing short-term motion estimation.

Challenges. Optical flow estimation remains an active and extensively studied research topic (Teed & Deng, 2020; Huang et al., 2022a; Sun et al., 2022; Dong & Fu, 2024). However, mainstream optical flow models trained on synthetic data with heavy augmentations often diverge significantly from the scenarios targeted by FFS. Moreover, the learning objectives of these models are not aligned with the goal of generating optical flow that facilitates accurate pixel warping and high-quality image synthesis. (Xue et al., 2019) point out that for different downstream tasks, it is often necessary to fine-tune or even train a flow estimation network from scratch. Interestingly, higher-performing optical flow networks may lead to degraded image synthesis results, as they often emphasize ambiguous regions such as occlusions and may lack sufficient spatial resolution (Niklaus & Liu, 2020; Huang et al., 2022b). In real-world scenarios, obtaining ground-truth optical flow labels remains a major challenge.

We believe that near-future frames can be synthesized with increasing accuracy as optical flow methods continue to improve. However, integrating optical flow methods into long-term video generation remains a significant challenge (Liang et al., 2024). Optical flow is typically limited to predicting pixel motion over very short time spans and cannot assist in generating novel video content.

3.1.4 Transformers.

Since the groundbreaking design of ViT (Dosovitskiy et al., 2021), which applied a pure transformer directly to sequences of image patches to extract continuous tokens, the use of transformers in frame synthesis has attracted significant attention. Video frame interpolation is a task that is closely related to FFS (Liu et al., 2017). Both (Shi et al., 2022) and (Lu et al., 2022) propose transformer-based video interpolation frameworks that overcome the limitations of traditional CNNs by leveraging self-attention mechanisms to capture long-range dependencies and enhance content awareness. These methods introduce innovative strategies, such as local attention in the spatiotemporal domain and cross-scale window-based attention, to improve performance and effectively handle large motions. (Ye & Bilodeau, 2023) presents an efficient transformer model for video prediction, leveraging a novel local spatiotemporal separation attention mechanism, and compares three variants—fully auto-regressive, partially auto-regressive, and non-autoregressive—to balance performance and complexity. Numerous ongoing studies continue to explore the use of transformers for modeling inter-frame dynamics under varying motion amplitudes and for addressing challenges in high-resolution frame synthesis (Park et al., 2023; Zhang et al., 2024).

Challenges. Many researchers believe that transformers perform particularly well on large-scale datasets (Zhai et al., 2022; Smith et al., 2023). However, most existing studies focus primarily on high-level vision tasks. Furthermore, the success of transformers in many tasks depends heavily on fully leveraging the capabilities of foundation models. In the context of image synthesis, best practices remain unclear (Li et al., 2023). Drawing insights from large language models (LLMs) and related architectures may offer a promising direction, which we further explore in Section 5.2.

3.2 Feature Space

Synthesizing in the raw pixel space often overburdens models, as it requires reconstructing images from scratch—a task especially challenging for high-resolution video datasets. This challenge has prompted a shift in focus among researchers. Rather than grappling with the complexities of pixel-level synthesis, many studies have shifted toward high-level feature synthesis in the feature space, focusing on representations such as segmentation maps and depth maps. These approaches provide a more efficient means of handling the complexities of video data (Oprea et al., 2020).

Future semantic segmentation. Future semantic segmentation represents a progressive approach to video synthesis, primarily focusing on generating semantic maps for future video frames. This methodology departs from traditional raw pixel forecasting by utilizing semantic maps to narrow the synthesis scope and enhance scene understanding. In this context, the S2S model (Luc et al., 2017) emerges as a pioneering end-to-end system. It processes RGB frames along with their corresponding semantic maps as both input and output. This integration not only advances future semantic segmentation but also enhances video frame prediction, demonstrating the advantages of semantic-level forecasting. Building on this foundation, SADM (Bei et al., 2021) introduces further innovation by integrating optical flow with semantic maps. This fusion leverages optical flow for motion tracking and semantic maps for appearance refinement, using the former to warp input frames and the latter to inpaint occluded regions.

Future depth prediction. Depth maps, as 2D data structures that encode 3D information, can provide models with enhanced perception of the 3D world at minimal computational cost. Predicting future depth maps can benefit FFS tasks. MAL (Liu et al., 2023) introduces a meta-learning framework with a two-branch architecture, comprising future depth prediction and an auxiliary image reconstruction task. This framework improves the quality of synthesized future frames, particularly in complex and dynamic scenes.

Challenges. Future prediction in the feature space presents significant challenges due to the complex interplay between temporal dynamics and spatial context. Models must capture intricate motion patterns and accurately predict depths or semantic regions, which requires a deep understanding of 3D scene structures and object interactions. Ensuring temporal consistency and spatial precision while handling occlusions, perspective changes, and complex backgrounds is crucial. The use of high-resolution feature maps and large-scale annotated datasets further increases computational and data demands. Generalizing to unseen scenes and objects remains a major challenge, requiring robust models capable of adapting to diverse visual appearances and contexts. These challenges underscore the need for innovative approaches, such as meta-auxiliary learning, to enhance future prediction capabilities.

4 Stochastic Synthesis

In its early stages, video synthesis was primarily regarded as a low-level computer vision task, with an emphasis on using deterministic algorithms to optimize pixel-level metrics such as MSE, PSNR, and SSIM. However, this approach inherently limits the creative potential of such models by constraining possible motion outcomes to a single, fixed trajectory (Oprea et al., 2020). In response to this limitation, the field of video synthesis has undergone a paradigm shift—from relying on short-term deterministic prediction to embracing long-term stochastic generation. This transition acknowledges that, although stochastic synthesis may produce results that deviate significantly from the ground truth, it is essential for fostering a deeper understanding and enhancing creativity in the modeling of video evolution.

In this section, we explore stochastic synthesis methods, including GANs and VAEs, which were originally categorized for their emphasis on modeling randomness and uncertainty in video motion. Although these models possess generative capabilities, they were originally more closely associated with stochasticity due to their primary focus on capturing the inherent variability and unpredictability of video sequences. In Section 5, we discuss generative synthesis approaches, such as diffusion models and auto-regressive models, which are explicitly designed to prioritize the generation of diverse, high-quality video content. This distinction reflects the evolving focus of video synthesis research, shifting from an emphasis on stochasticity to a broader emphasis on generative capability.

4.1 Stochasticity Modeling

Uncertain object motion can be modeled either by incorporating stochastic distributions into deterministic frameworks or by directly employing probabilistic models.

Stochastic distributions. In the early stages, VPN (Kalchbrenner et al., 2017) employs CNNs to perform multiple predictions in videos based on pixel distributions, while SV2P (Babaeizadeh et al., 2018) enhances an action-conditioned model (Finn et al., 2016) by introducing stochastic distribution estimation. Shifting the focus to a more holistic representation of video elements, the PFP model (Hu et al., 2020) proposes a

Method	Publication	Main Ideas
VPN (Kalchbrenner et al., 2017)	ICML'17	Model the temporal, spatial, and color structure of video tensors, and encode it as a four-dimensional dependency chain.
SV2P (Babaeizadeh et al., 2018)	ICLR'18	Propose a stochastic variational video prediction (SV2P) method, providing effective stochastic multi-frame prediction for real-world videos.
PFP (Hu et al., 2020)	ECCV'20	Introduce a conditional variational approach to model the stochasticity. Learn a representation that can be decoded to future segmentation, depth and optical flow.
SRVP (Franceschi et al., 2020)	ICML'20	Introduce a stochastic temporal model using a residual update rule in a latent space, inspired by differential equation discretization schemes.
PhyDNet (Guen & Thome, 2020)	CVPR'20	Disentangle PDE dynamics from unknown complementary information, and propose a recurrent physical cell to perform PDE-constrained prediction in latent space.
TPK (Walker et al., 2017)	ICCV'17	Decompose video prediction: Use a VAE to model future human poses, and a GAN to generate future frames conditioned on these poses.
SVG (Denton & Fergus, 2018)	ICML'18	Learn a prior model of uncertainty to generate frames by drawing samples, and combine them with a deterministic estimate to generate varied and sharp results.
Vid2Vid (Wang et al., 2018)	NeurIPS'18	Model a mapping function from a source video (e.g., segmentation) to a photorealistic video, using carefully-designed GAN and a spatiotemporal adversarial objective.
SAVP (Lee et al., 2018)	*preprint*	Combine the latent variational variable model and adversarially-trained models, to produce predictions that look more realistic and better cover the range of possible futures.
Retrospective Cycle GAN (Kwon & Park, 2019)	CVPR'19	Employ two discriminators to identify fake frames and fake contained image sequences. Predict future and past frames while enforcing the consistency of bi-directional prediction.
vRNN (Castrejon et al., 2019)	ICCV'19	Argue that the blurry predictions of VAEs may caused by underfitting, and suggest increasing expressiveness of the latent distributions and using higher capacity likelihood models.
GHVAE (Wu et al., 2021)	CVPR'21	Attempt to address the underfitting issue on large and diverse datasets, by greedily training each level of a hierarchical autoencoder to learn high-quality video predictions.
INR-V (Sen et al., 2022)	TMLR'22	Propose a video representation network that utilizes INRs and a meta-network to generate diverse novel videos, demonstrating superior performance in various video-based generative tasks
DIGAN (Yu et al., 2022)	ICLR'22	Introduce INRs into GAN to enhance motion dynamics, and utilize a motion discriminator that effectively detects unnatural motions.
StyleGAN-V (Skorokhodov et al., 2022)	CVPR'22	Extend the paradigm of neural representations to build a continuous-time generator, and design a holistic discriminator that aggregates temporal information by concatenating frames' features.
CDNA (Finn et al., 2016)	NeurIPS'16	Develop an action-conditioned video prediction model that explicitly models pixel motion. The model generalizes to unseen objects by decoupling motion and appearance.
AMC-GAN (Jang et al., 2018)	ICML'18	By providing appearance and motion information as conditions, reduce the prediction uncertainty of equally probable future.
MoCoGAN (Tulyakov et al., 2018)	CVPR'18	Generate a video by mapping a sequence of random vectors to a sequence of video frames, and introduce a novel adversarial learning scheme utilizing both image and video discriminators.
LMC (Lee et al., 2021)	CVPR'21	Study how to store abundant long-term contexts, and recall suitable motion context, especially complex motions from limited inputs.
SLAMP (Akan et al., 2021)	ICCV'21	Focus on learning stochastic variables for separate content and motion.
MMVP (Zhong et al., 2023)	ICCV'23	Construct appearance-agnostic motion matrices to decouple motion and appearance.
LEO (Wang et al., 2024c)	IJCV'24	Represent motion as a sequence of flow maps and synthesize videos in the pixel space, and utilize the latent motion diffusion model to model the motion distribution.
D-VDM (Shen et al., 2024)	AAAI'24	Decompose future frames into spatial content and temporal motions. Predict temporal motions based on a 3D-UNet diffusion model.
DrNet (Denton et al., 2017)	NeurIPS'17	Decompose video frames into stationary and time-varying components.
DVGPC (Cai et al., 2018)	ECCV'18	Tackle the severe ill-posedness of human action video prediction with a two-stage framework: generating a human pose sequence from random noise, then creating the human action video.
CVP (Ye et al., 2019)	ICCV'19	Predict the future states of independent entities while reasoning about their interactions, and then synthesize future frames.
OCVP-VP (Villar-Corrales et al., 2023)	ICIP'23	Decouple the processing of temporal dynamics and object interactions.
SlotFormer (Wu et al., 2023)	ICLR'23	Model spatio-temporal relationships by reasoning over object features, and predict accurate future states of objects.
OKID (Comas et al., 2023)	L4DC'23	Decompose a video into moving objects, their attributes and the dynamic modes.
MOSO (Sun et al., 2023)	CVPR'23	Identify motion, scene, and object as the pivotal elements of a video.

Table 3: Overview of stochastic synthesis methods.

probabilistic approach for simultaneously synthesizing semantic segmentation, depth maps, and optical flow. Additionally, SRVP (Franceschi et al., 2020) leverages Ordinary Differential Equations (ODEs), while PhyDNet (Guen & Thome, 2020) employs Partial Differential Equations (PDEs) to model stochastic dynamics. A potential drawback lies in their assumption that physical laws can be linearly disentangled from other factors of variation in the latent space—an assumption that may not hold for all types of videos.

Probabilistic models. Building on the pioneering work of Multi-Scale AdvGDL (Mathieu et al., 2016), adversarial training has substantially advanced FFS tasks by improving the prediction of uncertain object motions. Similarly, vRNN (Castrejon et al., 2019) and GHVAE (Wu et al., 2021) enhance VAEs through the incorporation of likelihood networks and hierarchical structures, respectively, thereby contributing a new dimension to the ongoing evolution of stochastic synthesis methods.

To address the challenges of pixel-level synthesis, several studies introduce intermediate representations. S2S (Luc et al., 2017) and Vid2Vid (Wang et al., 2018) incorporate adversarial training into future semantic segmentation frameworks. Additionally, the TPK model (Walker et al., 2017) leverages a VAE to extract human pose information, followed by a GAN to predict future poses and frames. It is worth noting that directly modeling stochastic distributions tends to yield broader predictive coverage but often results in poor visual quality. In contrast, probabilistic models can produce sharper results, but they often face challenges such as mode collapse, training instability, and high computational cost. Bridging these two methodologies, SAVP (Lee et al., 2018) integrates stochastic modeling with adversarial training, achieving a balance between broad predictive diversity and improved visual quality.

Recognizing that object motion is largely deterministic—except in cases of unforeseen events such as collisions, SVG (Denton & Fergus, 2018) models trajectory uncertainty using both fixed and learnable priors, effectively blending deterministic and probabilistic approaches. In a similar vein, but with an emphasis on temporal coherence, Retrospective Cycle GAN (Kwon & Park, 2019) introduces a sequence discriminator to detect fake frames. Building on the paradigm of implicit neural representations (INRs) for video (Sen et al., 2022), this concept of scrutinizing frame authenticity is further extended in DIGAN (Yu et al., 2022), where the focus shifts to a motion discriminator aimed at identifying unnatural motions. StyleGAN-V (Skorokhodov et al., 2022) highlights motion consistency from a different perspective by incorporating continuous motion representations into StyleGAN2 (Karras et al., 2020), enabling consistent generation in high-resolution settings.

Challenges. Although stochastic models are capable of capturing a broad spectrum of plausible futures, they often struggle with poor visual quality and increased computational demands. Direct modeling of stochastic distributions often leads to blurred outputs, whereas probabilistic models may encounter issues such as mode collapse and training instability. Striking a balance between diversity, visual fidelity, and computational efficiency remains a significant challenge. Moreover, the assumption that physical laws can be linearly disentangled from other factors of variation may not hold across all types of videos, underscoring the need for more adaptable and generalizable models.

4.2 Disentangling Components

Stochastic synthesis algorithms primarily focus on modeling the randomness inherent in motion. However, this focus often overlooks the processes of object emergence and disappearance in videos. As a result, many studies isolate motion from other video elements or artificially constrain its evolution, aiming to better understand motion dynamics while reducing the complexity of real-world scenarios.

Content and motion. Video synthesis algorithms address the inherent complexity of natural video sequences by emphasizing intricate visual details. To this end, they aim to model appearances through fine-grained local information while simultaneously capturing the dynamic global content of videos. However, in applications such as robotic navigation and autonomous driving, understanding object motion patterns takes precedence over visual fidelity. This shift in priorities has driven the development of algorithms that emphasize object motion prediction and the disentanglement of motion from appearance. A notable early work, CDNA (Finn et al., 2016), set a precedent by explicitly predicting object motion. It maintains appearance invariance, enabling generalization to unseen objects beyond the training set. MoCoGAN (Tulyakov et al., 2018) learns to disentangle motion from content in an unsupervised manner, and the use of separate encoders for content and motion has since been widely adopted in video prediction models. This concept is further explored in LMC (Lee et al., 2021), where the motion encoder predicts motion based on residual frames, and the content encoder extracts features from the input frame sequence. MMVP (Zhong et al., 2023) takes a different approach by using a single image encoder, followed by a two-stream network that separately handles motion prediction and appearance preservation before decoding. To address the stochastic nature of motion, AMC-GAN (Jang et al., 2018) models multiple plausible outcomes through adversarial training. In contrast, SLAMP (Akan et al., 2021) adopts a non-adversarial approach that focuses on learning stochastic variables for disentangled content and motion representations. Further advancing this line of research, LEO (Wang et al., 2024c) and D-VDM (Shen et al., 2024) leverage diffusion models to achieve more realistic content-motion disentanglement, demonstrating recent progress in this direction.

Foreground and background. In future frame prediction, the motion dynamics of foreground objects and background scenes often differ substantially. Foreground objects usually display more dynamic motion, while background scenes tend to remain relatively static. This distinction has motivated research to predict the motion of these components separately, enabling a more nuanced understanding of video dynamics. A notable contribution in this area is DrNet (Denton et al., 2017), which specifically targets scenarios where the background remains largely static across video frames. The model decomposes images into object content and pose, and leverages adversarial training to develop a scene discriminator that determines whether two pose vectors originate from the same video sequence. Similarly, OCVP-VP (Villar-Corrales et al., 2023) employs the slot-wise scene parsing network SAVi (Kipf et al., 2022) to segment scenes hierarchically from the scene level down to individual objects. By focusing on such videos, prediction models can streamline their learning process by eliminating the need to model complex scene dynamics. Both Human-centric tasks, such as predicting human movement and interaction with the environment, and object-centric tasks, such as tracking object motion and positioning, can benefit from this approach.

Human-centric. FFS often focuses on foreground motion, especially in scenarios involving complex human movements. A common assumption in these scenarios—reflected across many specialized datasets—is that the background remains relatively static, which is characteristic of datasets focusing on detailed human motion. This has led to a strong research focus on understanding and forecasting human poses to improve foreground motion prediction. A representative example is DVGPC (Cai et al., 2018), which predicts skeleton motion sequences and transforms them into pixel space using a skeleton-to-image transformer. This method effectively bridges abstract motion representations and the pixel-level demands of video prediction, demonstrating a nuanced understanding of the complexities inherent in human-centric FFS tasks.

Object-centric. The concept of object-centric video prediction was first introduced by CVP (Ye et al., 2019), laying the foundation for this specialized subfield of video prediction. SlotFormer (Wu et al., 2023) introduces transformer-based auto-regressive models to learn object-specific representations from video sequences. This design enables consistent and accurate tracking of individual objects over time. A more recent advancement, OKID (Comas et al., 2023), uniquely decomposes videos into distinct components—specifically, the attributes and trajectory dynamics of moving objects—by employing a Koopman operator. This approach offers a more granular method for analyzing object motion in video sequences, setting it apart from prior methods.

General. Methods focused on human poses or objects have shown considerable promise on specific video datasets, but their reliance on predefined structures and limited adaptability to dynamic backgrounds hinder generalization. This limitation is reflected in their performance: while effective under controlled conditions, they often falter when confronted with background variation, revealing insufficient versatility for broader applications. To bridge this gap, MOSO (Sun et al., 2023) proposes a unified framework that identifies motion, scene, and object as the three pivotal elements of a video. It further refines content analysis by distinguishing between scene and object—where the scene denotes the background and the object denotes the foreground—as a finer decomposition of video content. MOSO's core contribution is a two-stage network architecture designed for general-purpose video analysis. In the first stage, the MOSO-VQVAE model encodes video frames into token-level representations, trained via a video reconstruction task to learn informative embeddings. In the second stage, transformers are employed to handle masked token prediction, enhancing the model's temporal reasoning capabilities. This design enables the model to perform a variety of token-level tasks, including video prediction, interpolation, and unconditional video generation.

Challenges. Disentangling content from motion, or foreground from background, in videos is complex due to the intricate interplay between temporal dynamics and spatial context. Models must accurately capture and predict motion patterns, depth, and semantic regions, while maintaining temporal consistency and spatial accuracy. The presence of occlusions, perspective changes, and complex backgrounds further increases the difficulty. The use of high-resolution feature maps and large-scale annotated datasets further exacerbates computational demands. Generalizing to unseen scenes and objects remains a major challenge, requiring robust models capable of adapting to diverse visual appearances and contexts. Applying the concept of separate processing to the generative methods discussed later (in Section 5) also presents a significant challenge.

4.3 Motion-Controllable Synthesis

In the field of FFS, one specialized research direction has emerged that focuses on the explicit control of motion. This approach is distinct in its emphasis on forecasting future object positions based on user-defined instructions, in contrast to the conventional reliance on past motion trends. The central challenge in this domain lies in synthesizing videos that follow these direct instructions while preserving a natural and coherent flow—a task that demands a nuanced understanding of both user intent and motion dynamics within the video context. This challenge underscores the delicate balance between user control and automated imagination, signaling a significant shift in how FFS models are conceptualized and implemented.

Strokes. Since there is no historical motion information available for generating videos from one single still image, several methods have emerged that allow for interactive user control. iPOKE (Blattmann et al., 2021) introduces techniques in which local interactive strokes and pokes enable users to deform objects in a still image to generate a sequence of video frames. These strokes represent the user's intended motion for the objects. Building on this innovative direction, the Controllable-Cinemagraphs model (Mahapatra & Kulkarni, 2022) proposes a method for interactively controlling the animation of fluid elements. These advances underscore the growing importance of user-centric approaches in the domain of motion-controllable FFS.

Instructions. The integration of instructions across various modalities—including local strokes, sketches, and text—is becoming increasingly common in works aiming to capture user-specified motion trends. VideoComposer (Wang et al., 2023a) synthesizes videos by combining text descriptions, hand-drawn strokes, and sketches. This approach adheres to textual, spatial, and temporal constraints, leveraging latent video diffusion models and motion vectors to provide explicit dynamic guidance. Essentially, it can generate videos that align with user-defined motion strokes and shape sketches. In a similar vein, DragNUWA (Yin et al., 2023) primarily leverages text for content description and strokes for controlling future motion, enabling the generation of customizable videos. These approaches advance the field of video generation by broadening the spectrum of user input modalities.

Challenges. Achieving natural and coherent video synthesis under explicit user control remains a significant challenge. Models must accurately interpret user intent and generate videos that follow specified motion instructions while maintaining both temporal and spatial coherence. Striking a balance between user control and the model's autonomous imagination is essential. Ensuring that the generated videos are both visually compelling and contextually appropriate further increases the complexity, requiring a deep understanding of user intent and video dynamics.

5 Generative Synthesis

In video analysis, the focus shifts towards algorithms designed for generative video synthesis, especially when dealing with videos that exhibit stochastic emergence and disappearance processes. These events introduce unpredictability as objects spontaneously appear and disappear. These algorithms require a profound understanding of the underlying physical principles that govern the real world in order to tackle such complexities. Rather than relying on simplistic linear motion predictions extrapolated from historical frames, they address

Method	Publication	Main Ideas
MCVD (Voleti et al., 2022)	NeurIPS'22	Build models from simple non-recurrent 2D convolutional architectures, and train them by randomly masking all past or future frames.
Video LDM (Blattmann et al., 2023b)	CVPR'23	Introduce a temporal dimension into the latent space diffusion model, transforming an image generator into a video generator.
SEINE (Chen et al., 2024b)	ICLR'24	Propose a short-to-long video diffusion model, aimed at generating coherent long videos at the "story-level".
LFDM (Ni et al., 2023)	CVPR'23	Synthesize an optical flow sequence in the latent space based on given conditions to warp the input image.
Seer (Gu et al., 2024)	ICLR'24	Inflate a pretrained T2I model along the temporal axis, and integrate global sentence-level instructions into each generated frame.
Emu Video (Girdhar et al., 2024)	ECCV'24	Investigate key design aspects of text-to-image-to-video generative models, including noise scheduling for diffusion and multi-stage training.
DynamiCrafter (Xing et al., 2024)	ECCV'24	Incorporate image guidance into the text-to-video diffusion model, by projecting images into a text-aligned context space.
SparseCtrl (Guo et al., 2024)	ECCV'24	Keep the pre-trained T2V model unchanged and introduce an additional condition encoder to process input like sketches, depth maps, and RGB keyframes.
PEEKABOO (Jain et al., 2024)	CVPR'24	Integrate user-interactive control into the T2V model via a masked attention module, without requiring extra training or inference overhead.
MicroCinema (Wang et al., 2024b)	CVPR'24	First generate images, then use the Appearance Injection Network to enhance appearance preservation. Employ an Appearance Noise Prior to retain the capabilities of pre-trained models.
LivePhoto (Chen et al., 2024a)	ECCV'24	Enable users to control the temporal motion of images with text, and decode motion-related instructions into videos using a generator equipped with a motion module.
I2VGen-XL (Zhang et al., 2023)	*preprint*	Propose a cascaded approach to decouple the two factors of semantics and quality, enhancing details through optimization with high-quality data.
Art-v (Weng et al., 2024)	CVPRW'24	Generate frames conditioned on the previous one autoregressively. Maintain the capabilities of the pre-trained model while avoiding the need to model long-range motions.
GAIA-1 (Hu et al., 2023a)	*preprint*	Cast world modeling as an unsupervised sequence modeling problem by transforming text, video, and ego-vehicle actions into discrete tokens.
Video Transformer (Weissenborn et al., 2020)	ICLR'20	Study how pure transformer-based video classification models extract and encode long spatiotemporal token sequences, outperforming previous 3D CNNs.
LVT (Rakhimov et al., 2021)	VISIGRAPP'21	Reduce the computational requirements of video generation models by modeling dynamics in the latent space.
Nuwa (Wu et al., 2022a)	ECCV'22	A unified multimodal model that handles language, images, and videos, employing a 3D Nearby Attention mechanism to reduce computational complexity.
Nuwa-Infinity (Liang et al., 2022)	NeurIPS'22	For variable-length generation tasks targeting images of arbitrary sizes or long-duration videos, introduce a global patch-based and a local visual token-based auto-regressive model.
MMVG (Fu et al., 2023)	CVPR'23	Discretize video frames into visual tokens and propose a multimodal masked video generation approach to tackle the text-guided video completion task.
MAGVIT (Yu et al., 2023)	CVPR'23	Introduce a 3D tokenizer to quantize videos into spatiotemporal visual tokens, and propose an embedding strategy for masked video token modeling.
LVM (Bai et al., 2024)	CVPR'24	Represent diverse visual data as sequences of discrete tokens, and demonstrate that visual auto-regressive models can scale effectively.
MAGVIT-v2 (Yu et al., 2024)	ICLR'24	Explore the design of discrete token tokenizers to boost auto-regressive models, and introduce a lookup-free quantizer to enhance video tokenizers.
VideoPoet (Kondratyuk et al., 2024)	ICML'24	Incorporate multimodal generative objectives—including images, videos, text, and audio—within an auto-regressive transformer framework.

Table 4: Overview of generative synthesis methods.

the challenge using sophisticated and imaginative modeling techniques. As a result, tasks such as transforming one static image into one dynamic video—often referred to as the image animation problem—have emerged as promising candidates for applying generative video prediction techniques.

5.1 Diffusion-Based Generation

Diffusion models (Ho et al., 2020) have emerged as a dominant approach for image generation. Early attempts at video prediction (Ho et al., 2022; Yang et al., 2023; Harvey et al., 2022; Voleti et al., 2022; Singer et al., 2023), which utilize pixel-space diffusion models by extending the conventional U-Net (Ronneberger et al., 2015) architecture to 3D U-Net structures, are constrained to generating low-resolution and short video clips due to high computational demands. The Latent Diffusion Model (LDM) (Rombach et al., 2022) extends this capability into the latent space of images, significantly enhancing computational efficiency and reducing resource consumption. This advancement has paved the way for applying diffusion models to video generation (Blattmann et al., 2023b).

Latent diffusion model extensions. Extensions of LDM have demonstrated strong generative capabilities in video systhesis (Voleti et al., 2022). For instance, Video LDM (Blattmann et al., 2023b) leverages pre-trained image models to generate videos, enabling multi-modal, high-resolution, and long-term video synthesis. Similarly, SEINE (Chen et al., 2024b) introduces a versatile video diffusion model capable of generating transition sequences, thereby extending short clips into longer videos.

Text-guided video completion with additional information. Recent research efforts have focused on harnessing additional modalities alongside RGB images to accomplish the task of text-guided video completion. LFDM (Ni et al., 2023) extends latent diffusion models to synthesize optical flow sequences in the latent space based on textual guidance. Seer (Gu et al., 2024) inflates Stable Diffusion (Rombach et al., 2022) along the temporal axis, enabling the model to utilize natural language instructions and reference frames to envision multiple variations of future outcomes. Emu Video (Girdhar et al., 2024) generates an image conditioned on textual guidance and extrapolates it into a video, making it adaptable to diverse textual inputs. DynamiCrafter (Xing et al., 2024) extends text-guided image animation to open-domain image scenarios. SparseCtrl (Guo et al., 2024) supports sketch-to-video generation, depth-to-video generation, and video prediction with an expanded range of input modalities. Other methods, such as PEEKABOO (Jain et al., 2024), explore interactive synthesis, aiming to unlock unprecedented applications and creative potential.

Preserving text guidance. Some works aim to achieve a more precise interpretation of textual guidance and preserve this information across the temporal dimension. MicroCinema (Wang et al., 2024b) adopts a divide-and-conquer strategy to address challenges related to appearance and temporal coherence. It employs a two-stage generation pipeline, first creating an initial image using an existing text-to-image generator, and then introducing a dedicated text-guided video generation framework for motion modeling. LivePhoto (Chen et al., 2024a) proposes a framework that incorporates motion intensity as an auxiliary factor to enhance control over motion dynamics. It further introduces a text re-weighting mechanism to emphasize motion-related descriptions, demonstrating strong performance in text-guided video synthesis tasks. I2VGen-XL (Zhang et al., 2023) utilizes static images to provide semantic and quality-related guidance, highlighting the diversity of approaches in text-guided video synthesis.

Integrating autoregressive models. Despite the rising dominance of diffusion models in generative tasks, some researchers seek to preserve the architecture of auto-regressive models to simultaneously leverage the strengths of both paradigms (Weng et al., 2024). Early generative video synthesis algorithms were constrained by limited data availability and model scalability, yet they laid the groundwork for integrating LLMs and diffusion models in video generation. VDT (Lu et al., 2024) represents one of the earliest efforts to incorporate a transformer-based backbone into diffusion models for video generation. Inspired by the success of DiT (Peebles & Xie, 2023) in image synthesis, VDT replaces the conventional U-Net with a spatiotemporal transformer that explicitly models both spatial and temporal dependencies. W.A.L.T. (Gupta et al., 2025) provides the first successful empirical demonstration of a transformer-based backbone for jointly training image and video latent diffusion models. Recently, GAIA-1 (Hu et al., 2023a) and Sora (Brooks et al., 2024) have also leveraged the strengths of both diffusion models and LLMs to enable more creative, generalizable, and scalable video synthesis. RIVER (Davtyan et al., 2023) leverages flow matching (Lipman et al., 2023) for efficient video prediction by conditioning on a small set of past frames in the latent space of a pre-trained VQGAN. Collectively, these advances underscore the potential of diffusion models to produce high-quality, controllable, and diverse video content, thereby pushing the boundaries of what is achievable in computer vision and artificial intelligence.

Challenges. While diffusion models have made significant strides in video generation, several critical challenges remain. Ensuring temporal coherence and consistency across frames is essential for achieving realism, yet remains a major challenge (Chen et al., 2024c; Xu et al., 2024). The computational efficiency and scalability of diffusion models are hindered by their resource-intensive nature, which limits their broader adoption (Peebles & Xie, 2023). Issues of controllability and interpretability persist, as textual guidance does not always align with visual outcomes, and model behaviors often remain opaque. Data availability and diversity are critical for training robust models, but acquiring comprehensive and diverse datasets remains a major bottleneck.

5.2 Token-Based Generation

Diffusion-based methods have garnered significant attention in the realm of image and video generation. However, these models typically have smaller parameter scales compared to contemporary large-scale language models. Recent research has increasingly focused on exploring how LLMs can be employed for such tasks, leveraging their optimization techniques and accumulated insights to investigate the applicability of scaling laws in the visual domain.

Key components. Implementing transformer-based FFS requires two key components: an efficiently scalable LLM framework and a high-quality visual tokenizer. Innovations such as VQ-VAE (Van Den Oord et al., 2017) and VQGAN (Esser et al., 2021) integrate auto-regressive models with adversarial training strategies to tackle image quantization and tokenization. An effective visual tokenizer should minimize the number of tokens per image or video clip segment while preserving near-lossless visual fidelity. However, the substantial token requirements for lossless reconstruction—particularly for high-resolution images—present challenges in processing long video sequences during training, thereby limiting video generation capabilities.

Early transformer applications. Even before the advent of LLMs (Brown et al., 2020; Achiam et al., 2023), transformers had already made a significant impact on time-series modeling. Video Transformer (Weissenborn et al., 2020) pioneered the use of transformer architectures in video synthesis by introducing an auto-regressive modeling approach. Despite its success, it inherits common limitations of transformer-based models, including high training resource demands and slow inference speed.

Latent space modeling. The Latent Video Transformer (LVT) (Rakhimov et al., 2021) introduces a novel latent-space approach that autoregressively models temporal dynamics and predicts future features, significantly reducing computational overhead. Other works, such as VideoGen (Zhang et al., 2020b) and Video VQ-VAE (Walker et al., 2021), also leverage the VQ-VAE framework to extract discrete tokens for video prediction. In contrast, Phenaki (Villegas et al., 2022) improves upon the ViViT (Arnab et al., 2021) architecture to extract continuous tokens. The NUWA framework (Wu et al., 2022a) proposes a versatile 3D transformer-based encoder-decoder architecture that is adaptable to diverse data modalities and tasks, further demonstrating the potential of transformers in video synthesis. NUWA-Infinity (Liang et al., 2022) builds upon this with an innovative generation mechanism designed to enable infinite high-resolution video synthesis, reflecting ongoing efforts to unify generative tasks across modalities.

Sequential modeling and visual sentences. LVM (Bai et al., 2024) introduces sequential modeling to enhance the learning capacity of large-scale vision models, demonstrating the scalability and flexibility of sequence-based models in in-context learning. The concept of "visual sentence" is proposed, in which a sequence of intrinsically related images is organized analogously to a linguistic sentence. This allows the model to leverage sequential information for sentence continuation and other visual tasks without relying on non-pixel-level knowledge.

In-context learning in visual domain. In-context learning has also been explored within the visual domain. Painter (Wang et al., 2023b) proposes a general framework for visual learning that enables images to "speak" through in-context visual understanding, thereby enhancing both image generation and interpretation. SegGPT (Wang et al., 2023c) explores the use of a GPT-based architecture for image segmentation, introducing the concept of "segmenting everything" and demonstrating the potential for unified segmentation under unsupervised learning settings, thereby advancing generalization in visual segmentation tasks.

Advances in video generation. In the domain of video generation, MAGVIT (Yu et al., 2023) presents a masked generative video Transformer that efficiently processes videos by masking certain regions and predicting the missing segments. MAGVIT-v2 (Yu et al., 2024) suggests that transformer-based models may surpass diffusion models in visual generation tasks, highlighting the pivotal role of visual tokenizers. VideoPoet (Kondratyuk et al., 2024) introduces a large language model for zero-shot video generation, pushing the boundaries of unsupervised video synthesis. It enables users to generate or edit videos based on high-level textual prompts and excels at capturing temporal and contextual relationships within video data.

Text-guided generative video synthesis algorithms generate sequences of frames by integrating contextual visual information with textual guidance. The Text-guided Video Completion (TVC) task entails completing videos under various conditions, including the first frame (video prediction), the last frame (video rewind), or both (video transition), guided by textual instructions. MMVG (Fu et al., 2023) addresses the TVC task using an auto-regressive encoder-decoder architecture that integrates textual and visual features, forming a unified framework capable of handling diverse video synthesis tasks.

Collectively, these studies integrate visual tokenizers and large language models to establish unified and scalable frameworks for visual learning, thereby driving significant progress in FFS tasks. The evolving landscape of FFS research continues to highlight the potential of transformers and LLMs in unifying generative tasks across modalities.

Challenges. Token-based generation for FFS faces significant challenges, particularly in designing efficient visual tokenizers that balance a minimal number of tokens with near-lossless reconstruction, especially for high-resolution content. The high computational demands of transformer-based models also present barriers to adoption, particularly in resource-constrained environments. When computational resources are limited, the emergence phenomena reported in previous studies under this paradigm may fail to manifest (Bai et al., 2024). Recent work (Sun et al., 2024a) suggests that one reason token-based models struggle to match the visual quality of diffusion models is their limited integration with high-quality community assets, such as robust training infrastructure and curated datasets.

6 Application Realms

The applications of FFS span a wide range of domains, underscoring its growing significance across diverse fields.

World model. World models (Ha & Schmidhuber, 2018b;a; Zhu et al., 2024) provide a general-purpose framework for simulating and predicting the dynamics of complex systems. These models are widely employed in reinforcement learning and robotics, enabling agents to make informed decisions and perform actions that lead to desired outcomes. FFS serves as a key learning objective in the development of world models (Hafner et al., 2020; 2021; 2023; Wang et al., 2024a; Ge et al., 2024; Agarwal et al., 2025). Recent work (Escontrela et al., 2024) demonstrates that video prediction can also be incorporated into reward modeling to further support reinforcement learning. GameNGen (Valevski et al., 2025) has demonstrated exceptional world modeling performance and strong instruction-following ability in controllable future frame synthesis.

Autonomous driving. FFS is indispensable for autonomous vehicles, including self-driving cars and drones, as it enables them to anticipate the movement of objects, pedestrians, and other vehicles. This predictive capability is critical for ensuring safe and efficient navigation. For instance, GAIA-1 (Hu et al., 2023a) employs a unified world model that integrates multimodal large language models and diffusion processes to predict control signals and future frames, thereby enhancing decision-making capabilities in autonomous systems. In most existing driver-assistance systems, visual input is first transformed into structured representations—such as objects, lane markings, and traffic lights—followed by downstream predictions based on these modalities. Effectively leveraging raw visual information for trajectory prediction in real-world scenarios remains a significant challenge (Nayakanti et al., 2023; Varadarajan et al., 2022). Most existing trajectory prediction approaches rely solely on historical detection records of surrounding vehicles and pedestrians. Recent studies (Gu et al., 2023) have demonstrated that fully exploiting semantic information from visual inputs can significantly improve behavior prediction in dynamic scenes.

Robotics. In the field of robotics, future frame synthesis is employed to guide robotic agents through dynamic environments. It enables them to effectively plan paths, manipulate objects, and avoid obstacles, as demonstrated by (Finn & Levine, 2017). By predicting future states, robotic systems can make proactive decisions, thereby enhancing adaptability and operational efficiency in complex environments. The GR-1 and GR-2 approaches (Wu et al., 2024; Cheang et al., 2024) demonstrate that visual robotic manipulation can benefit significantly from large-scale video generative pretraining. After being pretrained on large-scale video datasets, GR-1/2 can be seamlessly fine-tuned on robot-specific data, exhibiting strong generalization to unseen scenes and objects.

Film production. Future frame synthesis has found valuable applications in the film industry, particularly in special effects, animation, and pre-visualization. It assists filmmakers in generating realistic scenes and enhancing the overall cinematic experience. For example, Mahapatra & Kulkarni (2022) utilize FFS to generate visually compelling sequences that enhance narrative coherence and support artistic expression in filmmaking.

Meteorological. Future frame synthesis (FFS) plays a vital role in weather forecasting by assisting meteorologists in simulating and predicting atmospheric dynamics. By accurately forecasting future spatiotemporal patterns, FFS enhances the precision of weather prediction models, as demonstrated by (Shi et al., 2017). This capability is essential for both operational weather forecasting and disaster preparedness.

Anomaly detection. (Liu et al., 2018) proposes a video anomaly detection approach based on future frame prediction, under the assumption that normal events are predictable whereas abnormal events deviate from expected patterns. The method introduces a motion constraint alongside appearance constraints to ensure that the predicted future frames align with the ground truth in both spatial and temporal dimensions.

Overall, these diverse applications underscore the significance and broad potential of future frame synthesis as a powerful tool for understanding and interacting with the dynamic world. Its capacity to predict future states from past observations positions it as a valuable asset across a wide spectrum of domains, including artificial intelligence, robotics, entertainment, and beyond.

Challenges. When applying FFS methods to real-world scenarios, several potential challenges may arise. For instance, domain-specific applications may require solutions for few-shot learning (Gui et al., 2018) or test-time adaptation (Choi et al., 2021). Another challenge lies in the effective integration of vectorized data with image-based inputs. Moreover, interpretability remains a concern in real-world applications, as end-to-end FFS methods often lack transparency and are difficult to interpret.

7 Related Work

Previous Surveys on Video Prediction. The field of video prediction, along with related areas such as action recognition and spatiotemporal predictive learning, has witnessed significant progress in recent years, largely driven by deep learning techniques. Several comprehensive surveys have provided overviews of state-of-the-art methods, benchmark datasets, and evaluation protocols in this domain. (Zhou et al., 2020) review next-frame prediction models developed prior to 2020, categorizing them into sequence-to-one and sequence-to-sequence architectures. The survey compares these approaches by analyzing their architectural designs and loss functions, and provides quantitative performance comparisons based on standard datasets and evaluation metrics. (Oprea et al., 2020) present a comprehensive review of deep learning methods for video prediction, outlining fundamental concepts and analyzing existing models based on a proposed taxonomy. The survey also includes experimental results to enable quantitative assessment of the state of the art. (Rasouli, 2020) provide an overview of vision-based prediction algorithms with a focus on deep learning approaches. They categorize prediction tasks into video prediction, action prediction, trajectory prediction, body motion prediction, and other related applications, and discuss common network architectures, training strategies, data modalities, evaluation metrics, and benchmark datasets. (Kong & Fu, 2022) survey state-of-the-art techniques in action recognition and prediction, covering existing models, representative algorithms, technical challenges, action datasets, evaluation protocols, and future research directions. (Tan et al., 2023) introduce OpenSTL, a unified benchmark for spatio-temporal predictive learning, which categorizes methods into recurrent-based and recurrent-free models. The paper provides standardized evaluations on multiple datasets and offers an in-depth analysis of how model architecture and dataset characteristics influence performance.

Surveys on Video Diffusion Models. (Xing et al., 2023) present a comprehensive review of video diffusion models in the era of AI-generated content, categorizing existing works into video generation, video editing, and other video understanding tasks. The survey provides an in-depth analysis of the literature in these areas and discusses current challenges and future research trends. (Li et al., 2024) survey recent advances in long video generation, summarizing existing methods into two key paradigms: divide-and-conquer and temporal auto-regressive modeling. They also provide a comprehensive overview and categorization of datasets and evaluation metrics, and discuss emerging challenges and future directions in this rapidly evolving field. (Sun et al., 2024b) review Sora, OpenAI's text-to-video model, categorizing the related literature into three themes—evolutionary generators, excellence in pursuit, and realistic panoramas—while also discussing datasets, evaluation metrics, existing challenges, and future directions. Complementarily, (Liu et al., 2024) provides a comprehensive analysis of Sora's underlying technologies, system design, current limitations, and its potential role in the broader landscape of large vision models.

Our Survey Focus. Our survey provides a comprehensive review of both historical and recent works in future frame synthesis, with a particular focus on the transition from deterministic to generative synthesis methodologies. The survey highlights key advances and methodological shifts, emphasizing the growing role of generative models in producing realistic and diverse future frame predictions.

8 Conclusion

In this survey, we have examined various aspects of future frame synthesis, including widely used datasets, evolving algorithmic paradigms, and prevailing challenges in the field.

Considering the broader development trends in artificial intelligence, we argue that the trajectory of video synthesis research should bifurcate. On the one hand, future research should focus on model lightweighting tailored for high-definition video applications, targeting low-level objectives such as video compression and short-term motion estimation. On the other hand, research should explore how models can develop a fundamental understanding of the physical world and generate content by leveraging substantial computational resources and diverse, long-duration video datasets. For the latter, future research should prioritize the development of evaluation metrics that incentivize stochastic synthesis, thereby expanding the potential for simulating the complexity of the human world. The ultimate objective is to develop video synthesis models with a profound understanding of the inherent dynamics within videos. Such models would be capable of generating videos over extended temporal horizons with high stochastic complexity in real-world scenarios.

Our proposed taxonomy is grounded in algorithmic stochasticity, highlighting the significant shift from deterministic approaches toward generative methodologies. This survey underscores the need to balance pixel-level accuracy with a deeper understanding of complex scene dynamics in video synthesis. Additionally, we examine the intricacies of stochastic emergence and disappearance processes, and advocate for improved evaluation metrics as well as the use of large-scale video datasets and substantial computational resources. We also categorize existing research directions and discuss the prevailing challenges in this domain. These insights are intended to inform and guide future research in video synthesis. As the field progresses, we anticipate the emergence of models with a more profound and nuanced understanding of real-world dynamics. Such advancements promise to improve accuracy, efficiency, and creative potential, paving the way for novel applications and future research.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016. URL `http://arxiv.org/abs/1609.08675`.

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

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Adil Kaan Akan, Erkut Erdem, Aykut Erdem, and Fatma Güney. Slamp: Stochastic latent appearance and motion prediction. In *ICCV*, 2021.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rk49Mg-CW`.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22861–22872, 2024.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Xinzhu Bei, Yanchao Yang, and Stefano Soatto. Learning semantic-aware dynamics for video prediction. In *CVPR*, 2021.

Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In *ICCV*, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023b.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

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.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *ECCV*, 2018.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *ICCV*, 2019.

Zhengping Che, Guangyu Li, Tracy Li, Bo Jiang, Xuefeng Shi, Xinsheng Zhang, Ying Lu, Guobin Wu, Yan Liu, and Jieping Ye. D2-city: a large-scale dashcam video dataset of diverse traffic scenarios. *arXiv preprint arXiv:1904.01975*, 2019.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Jiaben Chen and Huaizu Jiang. Sportsslomo: A new benchmark and baselines for human-centric video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6475–6486, 2024.

Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. Livephoto: Real image animation with text-guided motion control. In *European Conference on Computer Vision*, pp. 475–491. Springer, 2024a.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*, 2024b.

Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9232–9241, 2024c.

Xianhang Cheng and Zhenzhong Chen. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7029–7045, 2021.

Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10663–10671, 2020.

Myungsub Choi, Janghoon Choi, Sungyong Baik, Tae Hyun Kim, and Kyoung Mu Lee. Test-time adaptation for video frame interpolation via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9615–9628, 2021.

Armand Comas, Christian Fernandez Lopez, Sandesh Ghimire, Haolin Li, Mario Sznaier, and Octavia Camps. Learning object-centric dynamic modes from video and emerging properties. In *L4DC*, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In *CoRL*, 2019.

Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23263–23274, 2023.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *NeurIPS*, 2017.

Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34(4):743–761, 2011.

Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19068–19078, 2024.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NeurIPS*, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*. IEEE, 2017.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *NeurIPS*, 2016.

Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *ICML*, 2020.

Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *CVPR*, 2023.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *CVPR*, 2022.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. WorldGPT: Empowering LLM as multimodal world model. In *ACM Multimedia 2024*, 2024. URL `https://openreview.net/forum?id=G1tsqarGAw`.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 32(11):1231–1237, 2013.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In *ECCV*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *ECCV*, 2018.

Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5496–5506, 2023.

Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=qHGgNyQk31.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *CVPR*, 2020.

Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José MF Moura. Few-shot human motion prediction via meta-learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 432–450, 2018.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pp. 393–411. Springer, 2025.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018a.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018b.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

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

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic future prediction for video scene understanding. In *ECCV*, 2020.

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:2309.17080*, 2023a.

Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *CVPR*, 2023b.

Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, pp. 2439–2448, 2017.

Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European conference on computer vision*, pp. 668–685. Springer, 2022a.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, 2022b.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8079–8088, 2024.

Yunseok Jang, Gunhee Kim, and Yale Song. Video prediction with appearance and motion conditions. In *ICML*, 2018.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, 2013.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9000–9008, 2018.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *ICML*, 2017.

Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2071–2082, 2023.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33: 12104–12114, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *International Conference on Learning Representations (ICLR)*, 2022.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *Proceedings of Machine Learning Research*, 235:25105–25124, 2024.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*. IEEE, 2011.

Manoj Kumar, Neil Houlsby, Nal Kalchbrenner, and Ekin Dogus Cubuk. Do better imagenet classifiers assess perceptual similarity better? *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=qrGKGZZvH0`. Expert Certification.

Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *CVPR*, 2019.

Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv:1804.01523*, 2018.

Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *CVPR*, 2021.

Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. A survey on long video generation: Challenges, methods, and prospects. *arXiv preprint arXiv:2403.16407*, 2024.

Wenbo Li, Xin Lu, Shengju Qian, and Jiangbo Lu. On efficient transformer-based image pre-training for low-level vision. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 1089–1097, 2023.

Jian Liang, Chenfei Wu, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. In *NeurIPS*, 2022.

Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, pp. 56–74. Springer, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

Huan Liu, Zhixiang Chi, Yuanhao Yu, Yang Wang, Jun Chen, and Jin Tang. Meta-auxiliary learning for future depth prediction in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5756–5765, 2023.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, 2017.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=B1ewdt9xe`.

Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In *ICLR*, 2024.

Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3532–3542, 2022.

Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *ICCV*, 2017.

Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *CVPR*, 2022.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *4th International Conference on Learning Representations, ICLR 2016*, 2016.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2980–2987. IEEE, 2023.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, 2023.

Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5437–5446, 2020.

Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE international conference on computer vision*, pp. 261–270, 2017.

Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *TPAMI*, 44(6):2806–2826, 2020.

Junheum Park, Jintae Kim, and Chang-Su Kim. Biformer: Learning bilateral motion estimation via bilateral transformer for 4k video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1568–1577, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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:1704.00675*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

R Rakhimov, D Volkhonskiy, A Artemov, D Zorin, and E Burnaev. Latent video transformer. In *VISIGRAPP 2021-Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 101–112, 2021.

Amir Rasouli. Deep learning for vision-based prediction: A survey. *arXiv preprint arXiv:2007.00095*, 2020.

Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *ECCV*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016.

Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021.

Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *ICPR*. IEEE, 2004.

Bipasha Sen, Aditya Agarwal, Vinay P. Namboodiri, and C. V. Jawahar. INR-V: A continuous representation space for video-based generative tasks. *Trans. Mach. Learn. Res.*, 2022, 2022.

Cuifeng Shen, Yulu Gan, Chen Chen, Xiongwei Zhu, Lele Cheng, Tingting Gao, and Jinzhi Wang. Decouple content and motion for conditional image-to-video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4757–4765, 2024.

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. In *NeurIPS*, 2015.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *NeurIPS*, 2017.

Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17482–17491, 2022.

Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *ICCV*, 2021.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=nJfylDvgzlq`.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3626–3636, 2022.

Samuel L Smith, Andrew Brock, Leonard Berrada, and Soham De. Convnets match vision transformers at scale. *arXiv preprint arXiv:2310.16764*, 2023.

Li Song, Xun Tang, Wei Zhang, Xiaokang Yang, and Pingjian Xia. The sjtu 4k video sequence dataset. In *QoMEX*. IEEE, 2013.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015.

Deqing Sun, Charles Herrmann, Fitsum Reda, Michael Rubinstein, David J Fleet, and William T Freeman. Disentangling architecture and training for optical flow. In *European Conference on Computer Vision*, pp. 165–182. Springer, 2022.

Mingzhen Sun, Weining Wang, Xinxin Zhu, and Jing Liu. Moso: Decomposing motion, scene and object for video prediction. In *CVPR*, 2023.

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*, 2024a.

Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. From sora what we can see: A survey of text-to-video generation. *arXiv preprint arXiv:2405.10674*, 2024b.

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. *Advances in Neural Information Processing Systems*, 36:69819–69831, 2023.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 402–419. Springer, 2020.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59 (2):64–73, 2016.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=P8pqeEkn1H`.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.

Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 7814–7821. IEEE, 2022.

Angel Villar-Corrales, Ani Karapetyan, Andreas Boltres, and Sven Behnke. Mspred: Video prediction at multiple spatio-temporal scales with hierarchical recurrent networks. In *British Machine Vision Conference (BMVC)*, 2022.

Angel Villar-Corrales, Ismail Wahdan, and Sven Behnke. Object-centric video prediction via decoupling of object dynamics and interactions. In *Internation Conference on Image Processing (ICIP)*, 2023.

Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*, 2022.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, 2022.

Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017.

Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with VQVAE. *CoRR*, abs/2103.01950, 2021. URL `https://arxiv.org/abs/2103.01950`.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023a.

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*, 2024a.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023b.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1130–1140, 2023c.

Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, et al. Microcinema: A divide-and-conquer approach for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8414–8424, 2024b.

Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *IJCV*, 2024c.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024d.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *NeurIPS*, 2017.

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*, 2019. URL `https://openreview.net/forum?id=B1lKS2AqtX`.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rJgsskrFwH`.

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7395–7405, 2024.

Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *CVPR*, 2021.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022a.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NxoFmGgWC9`.

Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *CVPR*, 2020.

Yue Wu, Qiang Wen, and Qifeng Chen. Optimizing video prediction via video frame interpolation. In *CVPR*, 2022b.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=TFbwV6I0VLg`.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024.

Ziru Xu, Yunbo Wang, Mingsheng Long, Jianmin Wang, and M KLiss. Predcnn: Predictive learning with cascade convolutions. In *IJCAI*, 2018.

Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. *Advances in neural information processing systems*, 29, 2016.

Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 127:1106–1125, 2019.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14662–14672, 2024.

Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25(10):1469, 2023.

Xi Ye and Guillaume-Alexandre Bilodeau. Video prediction by efficient transformers. *Image and Vision Computing*, pp. 104612, 2023.

Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. Compositional video prediction. In *ICCV*, 2019.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv:2308.08089*, 2023.

Kihwan Yoon, Yong Han Kim, Sungjei Kim, and Jinwoo Jeong. Iam-vfi: Interpolate any motion for video frame interpolation with motion complexity map. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.

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 *CVPR*, 2023.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=gzqrANCF4g`.

Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=Czsdv-S4-w9`.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

Guozhen Zhang, Chunxu Liu, Yutao Cui, Xiaotong Zhao, Kai Ma, and Limin Wang. VFIMamba: Video frame interpolation with state space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=4s5UsBUsUS`.

Kai Zhang, Shuhang Gu, and Radu Timofte. Ntire 2020 challenge on perceptual extreme super-resolution: Methods and results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 492–493, 2020a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv:2311.04145*, 2023.

Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013.

Yunzhi Zhang, Wilson Yan, Pieter Abbeel, and Aravind Srinivas. Videogen: generative modeling of videos using vq-vae and transformers. 2020b.

Yiqi Zhong, Luming Liang, Ilya Zharkov, and Ulrich Neumann. Mmvp: Motion-matrix-based video prediction. In *ICCV*, 2023.

Yufan Zhou, Haiwei Dong, and Abdulmotaleb El Saddik. Deep learning in next-frame prediction: A benchmark review. *IEEE Access*, 8:69273–69283, 2020.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.