2024 IEEE International Automated Vehicle Validation Conference (IAVVC) | 979-8-3503-5407-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IAVVC63304.2024.10786438

DriveGenVLM: Real-world Video Generation for Vision Language Model based Autonomous Driving

Yongjie Fu, Anmol Jain, Xu Chen, Zhaobin Mo, and Xuan Di*

***Abstract*— The advancement of autonomous driving technologies necessitates increasingly sophisticated methods for understanding and predicting real-world scenarios. Vision language models (VLMs) are emerging as revolutionary tools with significant potential to influence autonomous driving. In this paper, we propose the DriveGenVLM framework to generate driving videos and use VLMs to understand them. To achieve this, we employ a video generation framework grounded in denoising diffusion probabilistic models (DDPM) aimed at predicting real-world video sequences. We then explore the adequacy of our generated videos for use in VLMs by employing a pre-trained model known as Efficient In-context Learning on Egocentric Videos (EILEV). The diffusion model is trained with the Waymo open dataset and evaluated using the Fréchet Video Distance (FVD) score to ensure the quality and realism of the generated videos. Corresponding narrations are provided by EILEV for these generated videos, which may be beneficial in the autonomous driving domain. These narrations can enhance traffic scene understanding, aid in navigation, and improve planning capabilities. The integration of video generation with VLMs in the DriveGenVLM framework represents a significant step forward in leveraging advanced AI models to address complex challenges in autonomous driving.**

I. INTRODUCTION

In the rapidly evolving field of autonomous driving, the integration of advanced predictive models into vehicular systems or transportation systems has become increasingly critical for enhancing safety and efficiency [1], [2]. Among the myriad of sensory technologies employed, camera-based video prediction stands out as a pivotal component, offering a dynamic and rich source of real-world data. Through the adoption of a cutting-edge diffusion model approach, this research not only contributes to the advancement of autonomous driving technologies but also sets a new benchmark for the application of predictive models in enhancing vehicular safety and navigational precision.

Content generated by artificial intelligence is presently a leading area of study within the domains of computer vision and artificial intelligence. The generation of photo-realistic and coherent videos is one of the challenging areas because of the limitations of memory and computation time. In the autonomous vehicle area, predicting the video from a vehicle's front camera is crucial for several reasons, particularly in the context of autonomous driving and advanced driver-assistance systems (ADAS) [3]. In this paper, we utilize the videos from the vehicle's surrounding cameras and predict future frames.

The generative model has also been utilized in the area of transportation and autonomous driving [4], [5]. Models are increasingly recognized for their capability to understand driving environments. Vision language models (VLMs) are now being utilized for autonomous driving applications. To enhance the utility of VLMs and explore the application of generative models to video content within VLMs, it is essential to validate generative models' predictions to confirm their relevance and accuracy in real-world scenarios. DriveGenVLM introduces the in-context VLM as a method to validate predicted videos from a diffusion-based generative model by providing textual descriptions of driving scenarios.

A. Related Work

Diffusion-based architectures have become increasingly popular in recent research for generating images and videos. Diffusion models have been applied to a variety of tasks for images, including image generation [6], image editing [7], and image-to-image translation [8]. Video generation and prediction are effective approaches to understand the real world. Several standard architectures have been utilized in this task, including Generative Adversarial Networks (GANs) [9], flow-based models, auto-regressive models, and Variational Autoencoders (VAEs) [10]. Recently, more diffusion models have been applied in this domain and achieve better video quality and more realistic frames, such as video generation [11] and text prompt to video generation [12].

Diffusion models are a class of deep generative models characterized by two main phases: (i) a forward diffusion phase, where the initial data is incrementally disturbed by the addition of Gaussian noise across multiple steps, and (ii) a reverse diffusion phase, where a generative model aims to reconstruct the original data from the noise-added version by progressively learning to invert the diffusion process, step by step. Denoising Diffusion Probabilistic Models (DDPM) represent a common type of generative model designed to learn and generate a specific target probability distribution through a diffusion process. DDPMs have been validated to be more effective than the traditional generation models, such as GANs and VAE.

*Corresponding author: Xuan Di.

‡This work is sponsored by NSF CPS-2038984 and NSF ERC-2133516.

Yongjie Fu is with the Department of Civil Engineering and Engineering Mechanics, Columbia University, New York, NY, 10027, USA (E-mail: yf2578@columbia.edu).

Anmol Jain is with the Department of Computer Science, Columbia University, New York, NY, 10027, USA (E-mail: aj3231@columbia.edu).

Xuan Di is with the Department of Civil Engineering and Engineering Mechanics, Columbia University, New York, NY, 10027 USA, and also with the Data Science Institute, Columbia University, New York, NY, 10027 USA (E-mail: sharon.di@columbia.edu).

Xu Chen and Zhaobin Mo are with the Department of Civil Engineering and Engineering Mechanics, Columbia University, New York, NY, 10027, USA (E-mail: xc2412@columbia.edu, zm2302@columbia.edu).

Generating long videos requires a large amount of computation sources. Some works overcome this challenge with autoregressive based models, such as Phenaki [12] and [13]. However, autoregressive models may lead to unrealistic scene transitions and persistent inconsistencies in extended video sequences because these models lack the opportunity to assimilate patterns from longer footage. To overcome this, MCVD [14] employs a training approach that prepares the model for various video generation tasks by independently and randomly masking either all preceding or subsequent frames. Meanwhile, FDM [11] introduces a framework based on Diffusion Probabilistic Models (DDPMs) that is capable of generating extended video sequences with realistic and coherent scene completion across diverse settings. NUWA-XL [15] introduces a "Diffusion over Diffusion" architecture designed for generating extended videos through a "coarse-to-fine" method.

In recent years, large language models (LLMs), which are text-based, have seen a surge in popularity [16], [17], [18]. Additionally, various generative vision-language models (VLMs) have been introduced in the autonomous driving domain. RAG-Driver [19] was proposed to leverage in-context learning for high performance, explainable autonomous driving. We leverage the in-context learning capabilities of EILEV [20] to generate descriptions of driving scenarios. In DriveGenVLM, the in-context VLMs allow us to process videos predicted by the diffusion framework, which can then be recognized by other vision-based models, potentially contributing to decision-making algorithms in autonomous driving. To the best of our knowledge, DriveGenVLM is the first work to integrate a video generation model and a Vision Language Model (VLM) into the autonomous driving domain.

B. Contributions of This Work

The key contributions of DriveGenVLM are summarized as follows:

1) Apply conditional denoising diffusion probabilistic models to the domain of driving video prediction.
2) Test the video generation framework in the Waymo open dataset of different camera angles to validate the feasibility for real world driving scenarios.
3) Utilize in-context vision language model to generate descriptions of the predicted video and validate that these videos can be applied for Vision language model based autonomous driving.

The rest of the paper is organized as follows. Sec. II introduces the preliminary knowledge used in this paper. Sec. III illustrates the solution approach. Sec. IV introduces the setting and results of the experiments. And Sec .V concludes this study.

II. Preliminory

A. Denoising Diffusion Probabilistic Models (DDPM)

The Denoising Diffusion Probabilistic Model is a type of generative model that has gained significant attention in the field of machine learning and computer vision [21]. DDPM operates through a forward process that transforms data into noise, and a backward process that reconstructs the original data from the noise. The goal of the forward process is to convert any data into a basic prior distribution, whereas the subsequent objective involves developing transition kernels to undo this conversion. To generate new data points, one begins by drawing a random vector from the prior distribution, then proceeds with ancestral sampling via the reverse Markov chain. The key to this sampling technique is to train the reverse Markov chain to replicate the time-reversed progression of the forward Markov chain accurately.

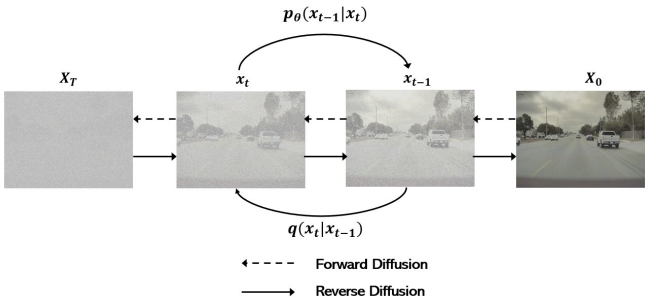

Fig. 1: Process of DDPM model.

For the conditional extension, in which the modeled x is conditioned on observations$y$. Given a data distribution $x_0 \sim q(x_0)$, the forward process generates a sequence of random variables $x_1, x_2, ..., x_T$. x_0 represents the original, noise-free data, while x_1 incorporates a slight amount of noise. This process continues up to x_T, which is nearly uncorrelated with x_0 and resembles a random sample drawn from a unit Gaussian distribution. The distribution of x_t depends only on x_{t-1}, the transition kernel is:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I). \tag{1}$$

The joint distribution is defined by the diffusion process and a data distribution $q(x_0, y)$ in Equ. 2.

$$q(x_{0:T}|y) = q(x_0, y)\prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2}$$

Denoting Diffusion Probabilistic Models (DPMs), these models operate by reversing the diffusion sequence. For given x_t and y, we use a neural network to estimate $p_\theta(x_{t-1}|x_t, y)$, serving as an approximation for $q(x_{t-1}|x_t, y)$. This estimation grants us the capability to procure samples of x_0 by commencing with the sampling of x_T from a standard Gaussian distribution, a choice made due to the diffusion process's initial state resembling a Gaussian distribution. Subsequently, we iteratively sample backwards, from x_T to x_0, through p_θ. The aggregate distribution of the sampled $x_{0:T}$ conditional on y is expressed as:

$$p_\theta(x_{0:T}|y) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y) \tag{3}$$

Here, $p(x_T)$ signifies a unit Gaussian distribution independent of θ. Training a conditional DPM entails the adjustment

of $p_\theta(x_{t-1}|x_t, y)$ to closely match $q(x_{t-1}|x_t, y)$ across the full range of t, x_t, and y values.

B. In-context Learning on VLMs

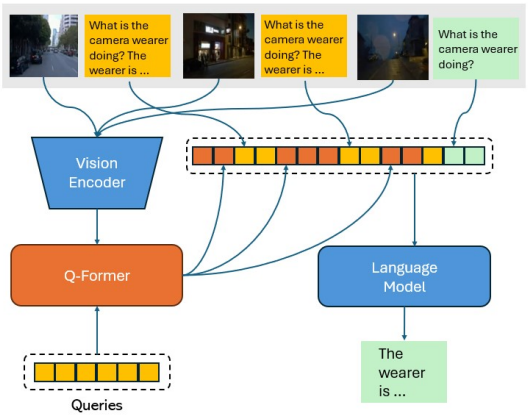

Fig. 3: Architecture of EILEV.

In-context learning was originally proposed in the paper of GPT-3 [22], which refers to the ability of a model to learn or adapt its responses based on the context provided within a single interaction, without any explicit updates or retraining on its underlying model.

We employ EILEV [20], a training technique developed to enhance in-context learning in Vision Language Models (VLMs) for first-person videos. As shown in Figure. 3, EILEV's architecture for an interleaved context-query scenario involves using the unmodified Vision transformer from BLIP-2 [23] to process video clips. The resulting compressed tokens are mixed with text tokens in the sequence of the initial context-query instance. These combined tokens are then input into BLIP-2's static language model, which produces new text tokens. This method can generalize out-of-distribution videos and texts and rare actions vis in-context learning. We make use of the pre-trained model to generate language narrations for the driving videos to validate that the generated results are explainable and realistic.

III. Methodology

Generating long, coherent, and photorealistic videos is still a challenge. The Flexible Diffusion Model (FDM) addresses this issue using a conditional generative model. We adopt a similar approach in DriveGenVLM. To sample coherent videos with a large number of frames, we can sample an arbitrary length of video condition on a small number of frames with a generative model. The goal is to sample coherent photo-realistic videos of driving scenarios with some frames. We utilize a sequential procedure to sample an arbitrarily long video with a generative model that can sample or condition on only a small number of frames at once.

Broadly, we define a sampling scheme as a series of tuples $[(X_s, Y_s)]_{s=1}^S$, where each tuple consists of a vector X_s indicating the indices of frames to be sampled and a vector Y_s showing the indices of frames to use as conditions for the stages $s = 1, \ldots, S$.

A. Training Architecture

We utilize a U-net structure for the DDPM image framework. This architecture is characterized by a sequence of layers that downscale spatial dimensions and then upscale them, interspersed with convolutional residual network blocks and layers that focus on spatial attention. The architecture is illustrated in Figure. 2. The DDPM iteratively transforms noise X_T to video frames X_0. The boxes with red borders are conditions. The right side shows the UNet architecture of each DDPM step.

Agorithm. 1 illustrates how we sample a video using a sample scheme. The generative model can sample any subsets of the video frames conditioned on other subsets. The model can generate any choice of X and Y.

Algorithm 1 Sample a video v given a sampling scheme $[(x_s, y_s)]_{s=1}^S$.

```
procedure SampleVideo(v; θ)
    for s ← 1, ..., S do
        y ← v[Y_s]
        x ~ DDPM(·; y, X_s, Y_s, θ)
        v[X_s] ← x
    end for
    return v
end procedure
```

B. Sampling Schemes

Sampling Schemes	Description
Autoreg	Samples x consecutive frames at each stage conditioned on the previous ten frames.
Hierarchy-2	Selects the first x frames (large groups) conditioned upon the previous ten frames, then samples consecutive frames within those groups until all frames are sampled.
Adaptive Hierarchy-2	Selects primary and secondary sampling frames, and adapts based on information gathered during the sampling process to optimize frame diversity using LPIPS distance.

TABLE I: Different Sampling Schemes.

Each sampling scheme's relative efficacy heavily relies on the dataset at hand, and there is no universally optimal option. In this work, we experimented with three sampling schemes as shown in Table I. The first and the most straightforward scheme adopted is the Autoreg, which samples ten consecutive frames at each step by conditioning on the previous ten frames. Another scheme used was Hierarchy-2 which employs a multi-tiered sampling approach, first tier with ten equidistantly chosen frames covering the unobserved portion of the video, conditioned upon ten observed frames. In the second tier, consecutive frames are sampled in groups, considering the nearest preceding and proceeding frames until all frames are sampled. Lastly, we used Adaptive Hierarchy-2 (Ad), which is only achievable through the implementation of FDM. Adaptive Hierarchy-2 strategically selects conditioning frames during testing to optimize frame

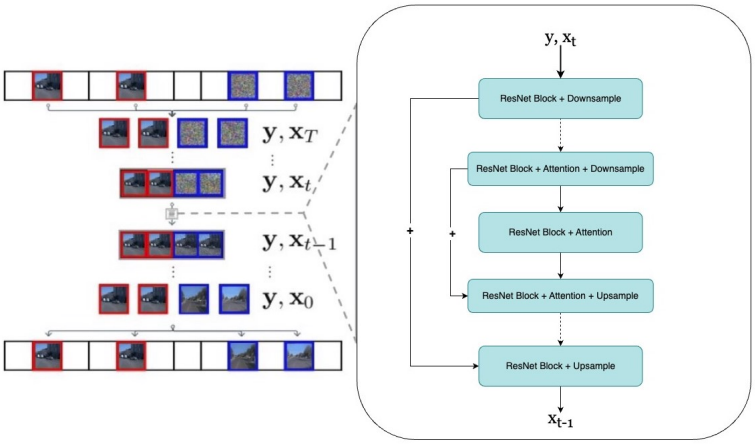

Fig. 2: Training Framework Employing U-Net with Diffusion Probabilistic Model (DDPM) Integration.

diversity, measured by the pairwise LPIPS distance between them.

IV. Experiments

A. Datasets

The Waymo Open Dataset [24] is a wide-ranging dataset that uses many sensors to aid in the progress of self-driving technology. It contains high-quality sensor data from Waymo's group of autonomous vehicles and is made up of more than 1,000 hours of videos. These videos are taken with various sensors such as LIDARs, radars, and five cameras (front and sides); they give a complete view around the car at all times or what we call 360-degree visibility. This group of data has very careful labeling, including marks for vehicles, people walking, bicycle riders and other things found on the road. This makes it extremely helpful for those working as researchers or engineers in this area to enhance their skills with perception (understanding), prediction (guessing what happens next) and simulation algorithms in self-driving cars. The Dataset V2 format is designed to be usable with Apache Parquet file formats and supported components. Here, a component is a set of related fields/columns that are required to understand each individual field.

B. Experiment Setup

To validate the algorithm in real-world driving scenarios, we utilize the Waymo Open Dataset, which encompasses diverse real-world environments across several cities. We extracted data for all the five present cameras in the dataset. We then pre-processed the datasets and extract the data from the three cameras being Front, Front-left, Front-right. In total we processed 138 videos. A total of 108 videos comprising of all three cameras divided equally were taken for training purposes, while the 10 videos for each of the three cameras for the test set. The maximum number of frames found for train videos was 199 frames, minimum contained around 175 frames. So we used 175 frames as the limit for all videos. The resolution was reduced to 128×128, and transformed into 4D tensors.

The model was operated on an 8-core Intel Cascade Lake processor and an NVIDIA L4 GPU with 24 GB memory in Debian GNU/Linux 11. A batch size of 1 with a learning rate of 0.0001 was used. The details of each camera training are shown in table II. The front was trained from scratch without using any pre-trained weights for $200,000$ iterations. Front-right used pre-trained weights from Camera-1 and was trained for $150,000$ iterations, and Front-left used pre-trained weights from Camera-3, trained for $100,000$ iterations. A total of 108 GPU hours were spent on training.

Camera Number	Camera Type	Iterations	GPU Hours
1	Front	200,000	48
3	Front-right	150,000	36
2	Front-left	100,000	24

TABLE II: Training details for each camera

C. Metrics

We utilize FVD (Fréchet Video Distance) [25], a metric used to evaluate the quality of videos generated by models in tasks like video generation or future frame prediction. Similar to the Fréchet Inception Distance (FID) used for images, FVD measures the similarity between the distribution of generated videos and real videos. FVD is useful for assessing the temporal coherence and visual quality of videos, making it a valuable tool for benchmarking video synthesis models.

D. Results

The FVD scores from our experiments on the Waymo Open Dataset for three cameras, which are tested using various sampling schemes, are summarized in Tables III IV V. The adaptive hierarchy-2 sampling method outperforms the other two methods.

Sampling Scheme	FVD Score
hierarchy-2	1489.0138
autoreg	1266.354
adaptive hierarchy-2	1174.563

TABLE III: FVD Scores for Front Camera.

Sampling Scheme	FVD Score
hierarchy-2	1295.744
autoreg	1401.793
adaptive hierarchy-2	812.425

TABLE IV: FVD Scores for Front-left Camera.

Sampling Scheme	FVD Score
hierarchy-2	1214.684
autoreg	1338.234
adaptive hierarchy-2	1122.159

TABLE V: FVD Scores for Front-right Camera.

Fig. 4: Front Camera - FVD Score: 1174.

Fig. 5: Front-left Camera - FVD Score: 812.

Figure. 4 - 6 shows prediction videos generated for each of the three cameras, using the Adaptive Hierarchy-2 sampling schemes yielding the lowest FVD scores. Each sub-figure contains 2 examples of generation videos for each camera. The frames with red bounding boxes are the ground truth frames, and the predicted frames are below each corresponding frame. The generated videos were conditioned on the first 40 frames for each example.

Fig. 6: Front-right Camera - FVD Score: 1122.

The FDM's training on the Waymo dataset showcased its capacity for coherence and photorealism. However, it still struggles with accurately interpreting the complex logic of real-world driving, such as navigating traffic and pedestrians. This limitation is likely due to the additional challenges present in real-world scenarios, which are absent in simulated environments.

E. Prediction Validation by in-context learning.

To validate that our generated videos are explainable and usable in vision language models, we employ the EILEV pre-trained model on Ego4D, eilev-blip2-opt-2.7b [20] to test our generated driving videos.

We utilize video clips and text pairs that describe the camera angle, driving environment, and time of day. The results are illustrated in Figure. 7. The action narrations generated by the model are displayed in an orange box. Notably, none of the verb and noun class combinations are shared in the first two videos, as shown in the blue box. As we can observe, the model can identify that the vehicle is driving on a highway with the camera positioned at the front. For the second video, the model recognizes that the vehicle is driving at night with its front camera. The in-context learning pre-trained model on VLMs performs well with the generated model, indicating that the videos are explainable and potentially usable by VLMs-based algorithms.

V. Conclusions

In summary, training the Denoising Diffusion Probabilistic Model (DDPM) on the Waymo dataset has shown its capability to produce coherent and lifelike images from both front and side cameras. However, it continues to face challenges in accurately capturing the complex dynamics of real-world driving, such as detailing buildings and tracking pedestrian movements. These difficulties are likely due to the complexities inherent in actual driving conditions, which are absent in synthetic datasets.

To explore potential applications of generated videos in Vision-Language Models (VLMs) for autonomous driving, we utilize the pre-trained EILEV model, an in-context VLM, to generate action narrations for the videos. The results indicate that the model can recognize unseen scenarios

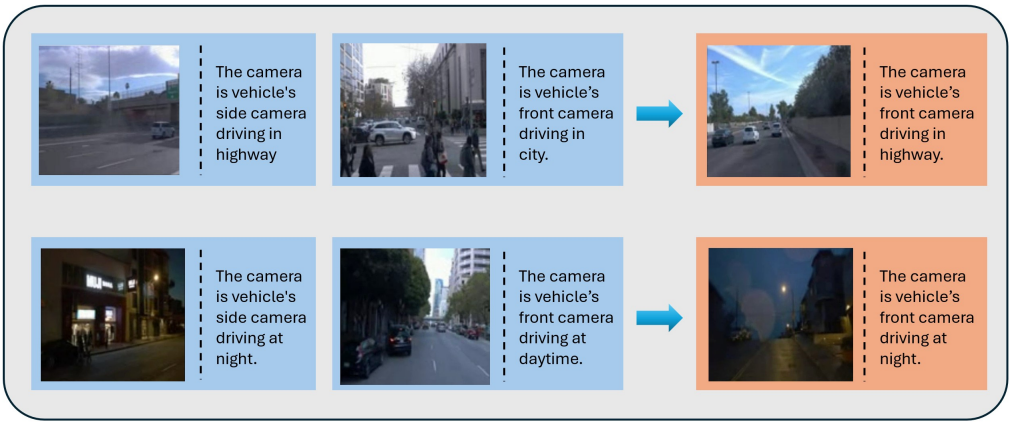

Fig. 7: The VLM model with in-context learning capabilities can generate action narrations for unseen driving videos.

and generate accurate narrations, demonstrating the potential for deploying VLM-based autonomous driving models that leverage outputs from generative models. The DriveGenVLM framework highlights the potential for using generative models and Vision Language Models (VLMs) together in autonomous driving tasks. For downstream applications, once we obtain narrations of driving scenarios, we can employ large language models to provide guidance to the driver or some language model-based algorithms.

REFERENCES

[1] Y. Fu, M. K. Turkcan, V. Anantha, Z. Kostic, G. Zussman, and X. Di, "Digital twin for pedestrian safety warning at a single urban traffic intersection," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 2640–2645.

[2] Y. Fu and X. Di, "Federated reinforcement learning for adaptive traffic signal control: A case study in new york city," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 5738–5743.

[3] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems," *Computer Vision and Image Understanding*, vol. 134, pp. 130–140, 2015.

[4] Z. Mo, Y. Fu, D. Xu, and X. Di, "Trafficflowgan: Physics-informed flow based generative adversarial network for uncertainty quantification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 323–339.

[5] Z. Mo, Y. Fu, and X. Di, "Quantifying uncertainty in traffic state estimation using generative adversarial networks," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 2769–2774.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[7] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

[8] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Vqbb: Image-to-image translation with vector quantized brownian bridge," *arXiv preprint arXiv:2205.07680*, 2022.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood, "Flexible diffusion modeling of long videos," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 953–27 965, 2022.

[12] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual description," *arXiv preprint arXiv:2210.02399*, 2022.

[13] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh, "Long video generation with time-agnostic vqgan and time-sensitive transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 102–118.

[14] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Mcvd-masked conditional video diffusion for prediction, generation, and interpolation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 371–23 385, 2022.

[15] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni, Z. Yang, L. Li, S. Liu, F. Yang, *et al.*, "Nuwa-xl: Diffusion over diffusion for extremely long video generation," *arXiv preprint arXiv:2303.12346*, 2023.

[16] B. Wang, H. Duan, Y. Feng, X. Chen, Y. Fu, Z. Mo, and X. Di, "Can llms understand social norms in autonomous driving games?" 2024. [Online]. Available: https://arxiv.org/abs/2408.12680

[17] G. Bai, Z. Chai, C. Ling, S. Wang, J. Lu, N. Zhang, T. Shi, Z. Yu, M. Zhu, Y. Zhang, *et al.*, "Beyond efficiency: A systematic survey of resource-efficient large language models," *arXiv preprint arXiv:2401.00625*, 2024.

[18] G. Bai, Y. Li, C. Ling, K. Kim, and L. Zhao, "Gradient-free adaptive global pruning for pre-trained language models," *arXiv preprint arXiv:2402.17946*, 2024.

[19] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, and M. Gadd, "Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," *arXiv preprint arXiv:2402.10828*, 2024.

[20] K. P. Yu, Z. Zhang, F. Hu, and J. Chai, "Efficient in-context learning in vision-language models for egocentric videos," *arXiv preprint arXiv:2311.17041*, 2023.

[21] Y. Fu, Y. Li, and X. Di, "Gendds: Generating diverse driving video scenarios with prompt-to-video generative model," 2024. [Online]. Available: https://arxiv.org/abs/2408.15868

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[23] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[24] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[25] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.