Denoising Diffusion Models for Dynamic Sky Image Generation

Sviatoslav Gladkykh, Supervisor: doc. RNDr. Elena Šikudová, Ph.D.

Faculty of Mathematics and Physics, Charles University

Introduction

Generative AI has made significant strides with the development of denoising diffusion models (DDMs), which excel in image generation tasks. This work focuses on applying DDMs to create dynamic sky images using a dataset of 360° sky images from the Computer Graphics Group (CGG). By investigating various DDM variants, including conditional and video diffusion models, this research aims to generate high-quality static and video sequences of sky images, enhancing their application in realistic image and video prediction.

Models and Methods

- Denoising Diffusion Probabilistic Models (DDPMs) were used to generate high-fidelity, visually realistic sky images.
- Denoising Diffusion Implicit Models (DDIMs) were employed to enhance image generation with faster sampling.
- Conditional Diffusion Models were utilized to predict the next frame in a sequence to capture dynamic changes in the sky.
- Video Diffusion Models (VDM) were applied to generate realistic video sequences of the sky.
- Random-Mask Video Diffusion Models (RaMViD) were used for video prediction, infilling, and upsampling.

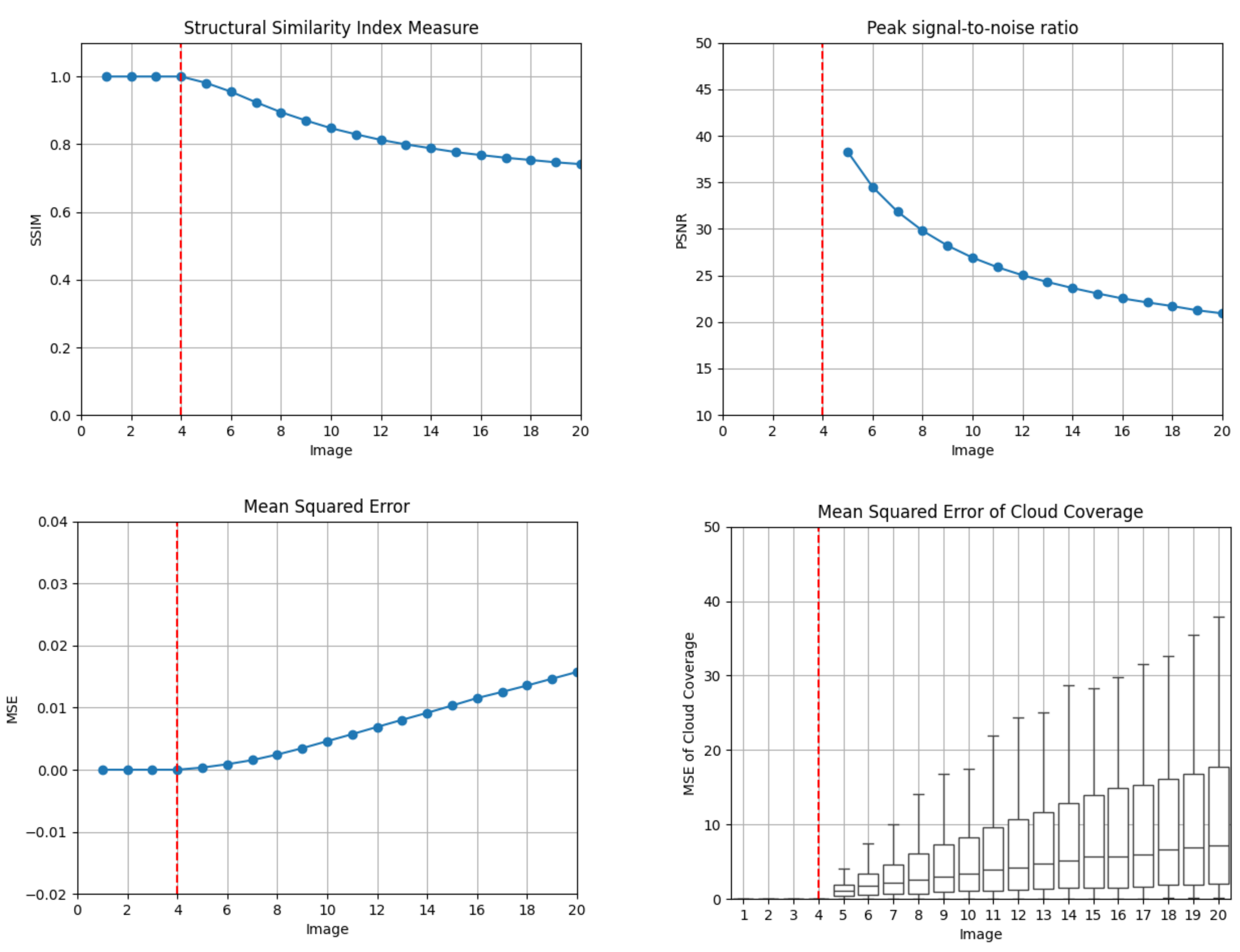
Results

Unconditional Image Generation
By implementing DDPMs and implicit models with tailored U-Net architectures featuring advanced components like residual blocks and attention mechanisms, we attained substantial improvements in image quality. Our model achieved an impressive Fréchet Inception Distance (FID) score of 5.43 after 300 epochs, showcasing superior performance compared to previous approaches.

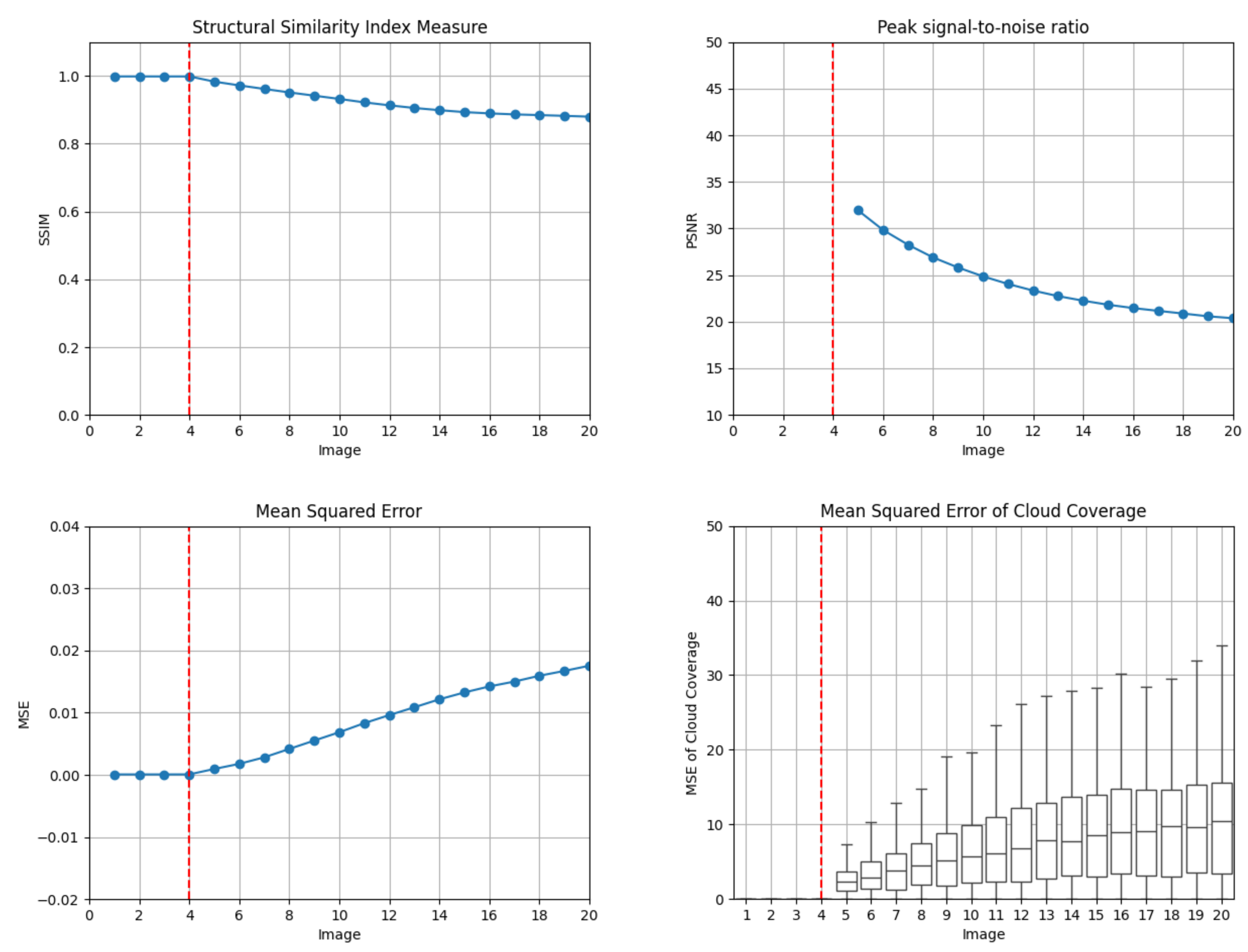
Conditional Image Generation
The outcomes leveraging concatenated conditioning frames demonstrate effective conditional image generation using denoising diffusion models. The results validate the effectiveness of the approach in maintaining stable cloud coverage estimates and producing visually coherent sequences.

Unconditional Video Generation
The research extended diffusion models to the 3D case for video generation using a model based on a 3D U-Net architecture. It successfully generated short videos of the sky and achieved a Fréchet video distance score of 154.7. This score indicates high-quality video generation, demonstrating the model's ability to capture both spatial and temporal details effectively.

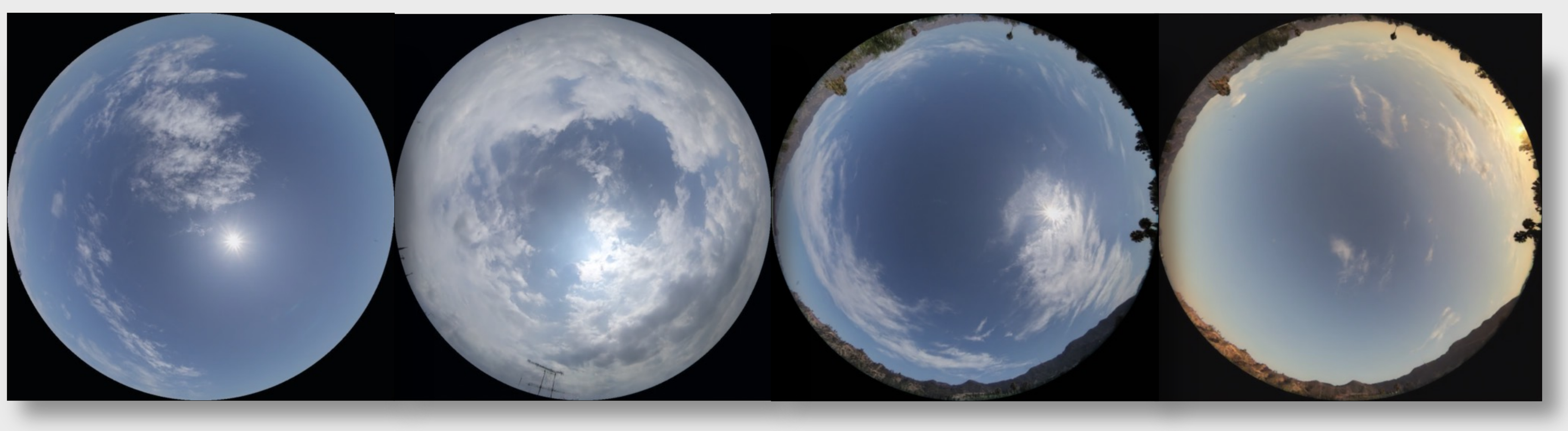
Conditional Video Generation
The random-mask video diffusion models, leveraging a novel conditioning technique during training, demonstrated improved performance in generating video sequences. This approach led to advancements in quality metrics and convergence rates.



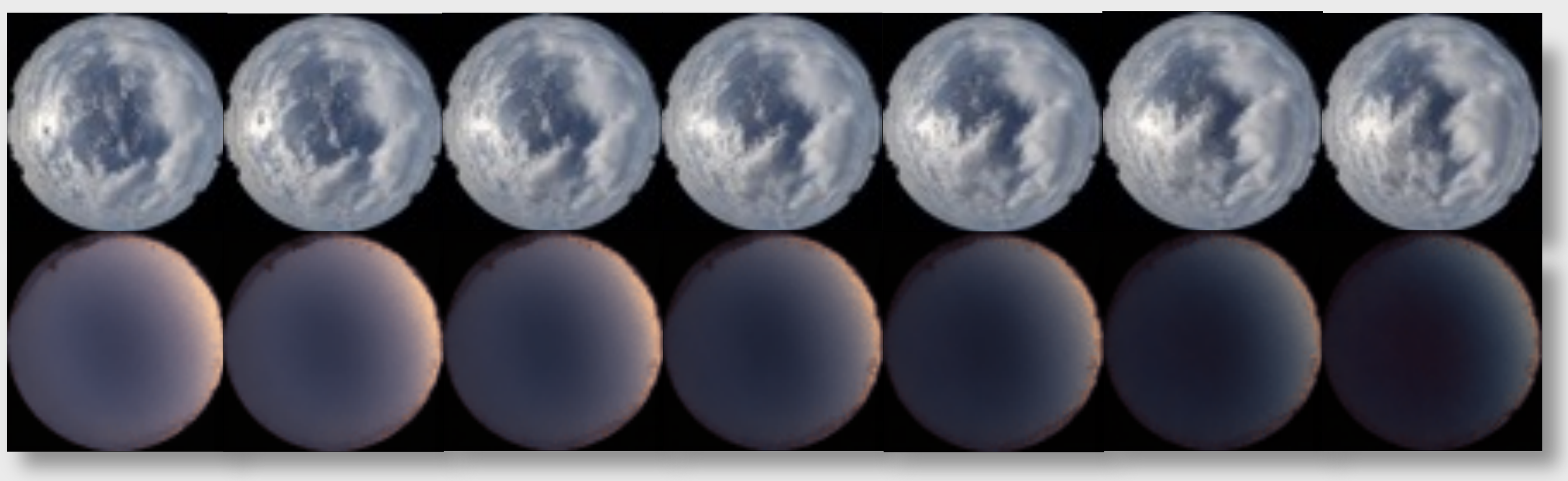
Charts showing quality evaluation metrics for videos generated by conditional DDPM



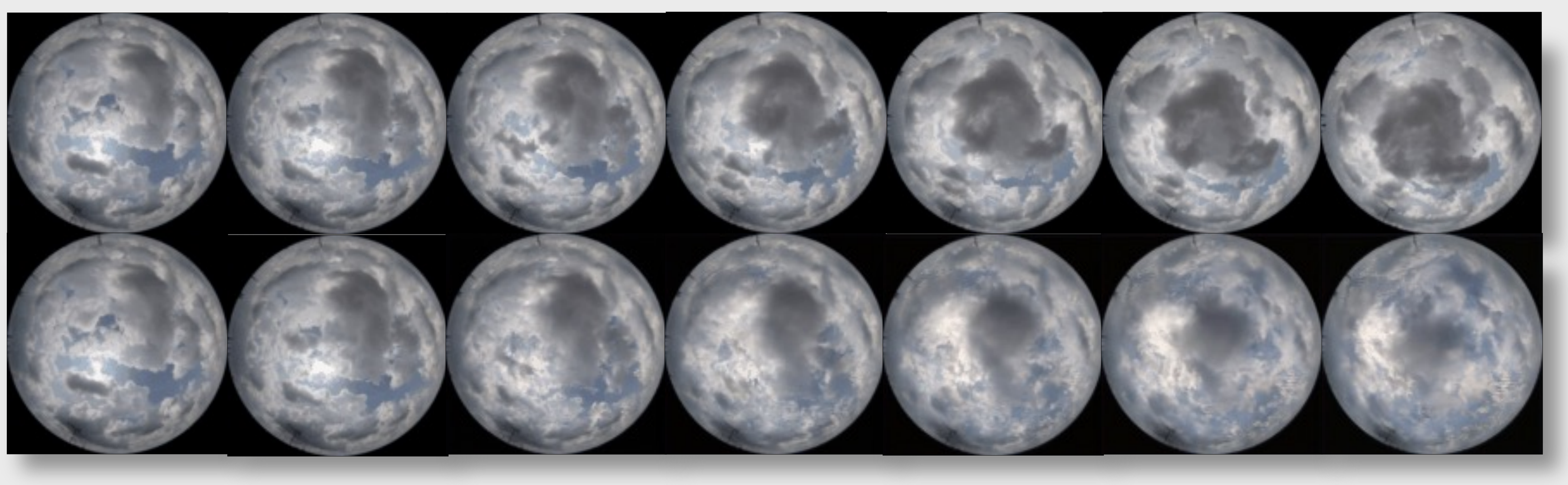
Charts showing quality evaluation metrics for videos generated by RaMViD



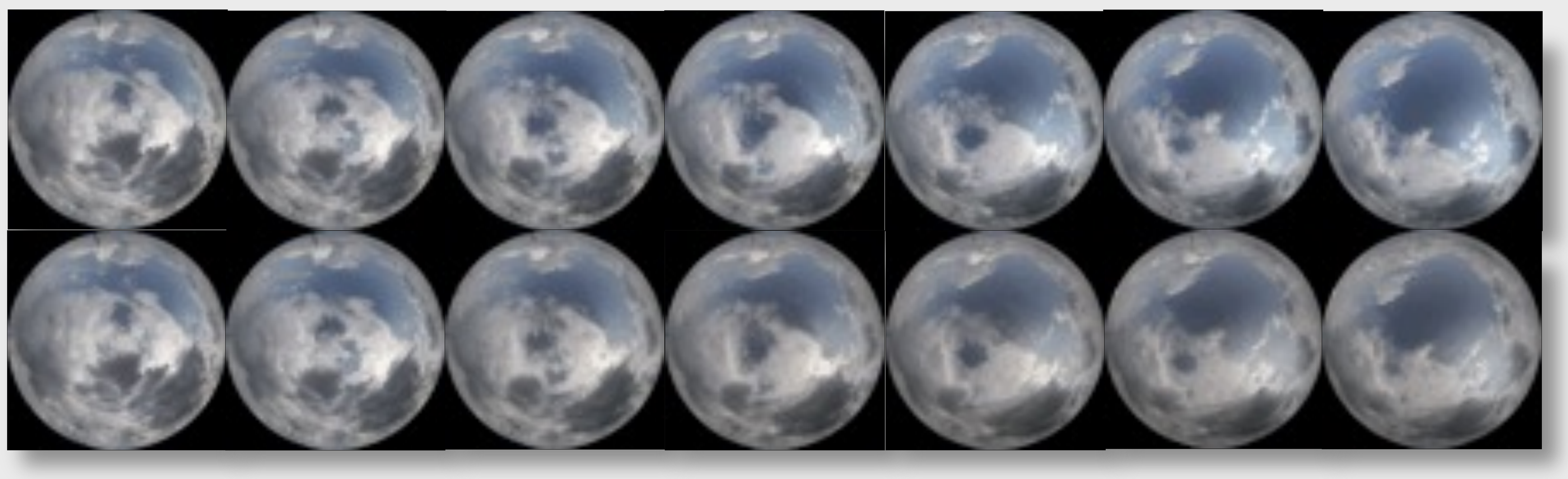
Samples of unconditionally generated images produced by DDPM



Samples of two sequences of images extracted from a 20-frame video, unconditionally generated by VDM



Two rows of 7 frames are taken from a 20-frame video. The first row presents the original frames, and the second row shows frames predicted by conditional DDPM. The first two frames in each row match, providing the model's conditioning



The two rows each have 7 frames from a 20-frame video. The top row shows original video frames, and the bottom row displays frames predicted by conditional DDPM. The first two frames in each row are identical, used as conditioning for the model

Conclusions

The comprehensive study demonstrates that diffusion models significantly enhance the quality of both unconditional and conditional image and video generation. They consistently outperform traditional GANs, producing highly realistic and accurate visuals, particularly for complex tasks like predicting sky conditions. This indicates diffusion models' superior potential for generating high-fidelity visual content across various applications.

Future work

Future work will focus on enhancing model capabilities for higher-resolution image and video generation, potentially through super-resolution techniques. Additionally, exploring different conditioning strategies and incorporating environmental data (e.g., time, location) could improve predictive accuracy and visual quality. Extended training and architecture optimization are also promising paths to fully realize the potential of diffusion models, especially for complex and long-term visual predictions.