Senior Electrical Engineering Project 2102499 Year 2025

SkyGPT revisited

Mr. Tiranut Kanjanatunyarut 6530141421

Mr. Wachirawit Piyaprapapan 6530349721

Advisor Dr. Suwichaya Suwanwimolkul

Department of Electrical Engineering, Faculty of Engineering
Chulalongkorn University
Academic Year 2025

Advisor signature	Co-advisor signature	Industry representative
Suwichaya		
(Suwichaya Suwanwimolkul)	(.....)	(.....)
Date 24 April 2026.	Date	Date

ABSTRACT

Solar photovoltaic (PV) power generation is directly dependent on solar irradiance, where the stochastic movement of clouds causes significant instability in power output. This volatility presents challenges for grid stability, necessitating advanced ultra-short-term forecasting methods to manage uncertainty. This senior project evaluates and optimizes the SkyGPT framework, a physics-informed generative model designed for probabilistic solar forecasting. While initial evaluations of the baseline model indicated suboptimal reconstruction quality, this study demonstrates that these limitations were primarily due to insufficient latent capacity and training instability rather than inherent flaws in the discrete latent modeling approach.

Methodologically, the study adopts a multi-stage, multi-stage workflow. The primary focus is the systematic optimization of the Vector-Quantized Variational Autoencoder (VQ-VAE) module. Results indicate that a larger codebook, deeper residual layers, and stabilized quantization via Exponential Moving Averages (EMA) are critical for preserving cloud boundaries and motion textures. The optimized model achieved a validation Peak Signal-to-Noise Ratio (PSNR) of 37.32 dB, significantly outperforming earlier baseline configurations. Integration into the full SkyGPT pipeline demonstrates stable temporal forecasting with a validation PSNR of 30.53 dB. Qualitative analysis reveals that while the framework can at capturing complex dynamics in partly cloudy conditions, performance limitations persist in overcast scenarios due to pixel-wise loss averaging. This work contributes a rigorous, reproducible benchmark that clarifies the architectural requirements for high-fidelity cloud-motion modeling.

Keywords: Solar photovoltaic, SkyGPT, Vector-Quantized Variational Autoencoder (VQ-VAE), Probabilistic Forecasting, Cloud-motion modeling.

AI Disclosure: Generative AI tools were utilized in the preparation of this report to refine language, ensure grammatical accuracy, and improve the professional tone of the text. All technical findings, experimental results, and core research conclusions remain the original work of the authors.

Contents

1 Introduction **7**

1.1 Background and Motivation . . . 7
1.2 Problem Statement . . . 7
1.3 Objectives and Contributions . . . 8
1.4 Scope of Work . . . 8
1.5 Expected Outcomes . . . 9

2 Literature review **10**

2.1 Theoretical Foundations for Generative Compression . . . 10
2.1.1 Overview of the SkyGPT . . . 10
2.1.2 Variational Autoencoders and Quantization . . . 11
2.1.3 VQ-VAE's loss function . . . 12
2.2 Temporal Prediction in Discrete Latent Space . . . 12
2.2.1 Autoregressive Sequence Modeling . . . 12
2.2.2 Frame Conditioning and Physical Constraints . . . 13
2.3 Evaluation Criteria . . . 13
2.4 Related Work . . . 13
2.4.1 Diffusion Models for High-Resolution Solar Forecasts . . . 13
2.4.2 Multimodal for Probabilistic Solar Forecasting . . . 14
2.4.3 Diffusion-Based Multi-Resolution Solar Power Prediction . . . 14
2.4.4 Video Representation Learning . . . 15
2.4.5 Discrete Representation Learning for Domain-Specific Data . . . 15
2.5 Evaluation Metrics . . . 16

3 Methodology **17**

3.1 Research Workflow . . . 17
3.2 Dataset Description . . . 17
3.3 Experimental Design . . . 18
3.3.1 Modified loss function in VQ-VAE . . . 18
3.3.2 Codebook Stability, Perplexity, and Latent Capacity . . . 18
3.3.3 Hyperparameter Search Strategy . . . 20
3.3.4 Full Pipeline Integration . . . 20
3.4 VQ-VAE Architecture and Configuration . . . 20
3.4.1 Encoder-Decoder Structure . . . 20
3.4.2 Loss Function and Optimization Strategy . . . 21
3.5 SkyGPT Temporal Prediction Context . . . 22
3.6 New Sky Image Dataset at CUEE . . . 22

4 Results and discussion **24**
4.1 Overview of Benchmark Evidence . . . 24
4.2 Top Performing Model Configurations . . . 24
4.3 Sensitivity Analysis . . . 24
4.3.1 Architectural Capacity . . . 24
4.3.2 Loss Balancing and Optimization . . . 25
4.3.3 Dataset Differences . . . 25
4.4 Qualitative Interpretation of the VQ-VAE Problem . . . 25
4.5 Figures for Benchmark Review . . . 26
4.6 SkyGPT Training Dynamics and Reconstruction Performance . . . 27
4.7 SkyGPT Reconstruction Performance . . . 28
4.7.1 Clear Sky Conditions . . . 28
4.7.2 Partly Cloudy and Dynamic Cloud Motion . . . 29
4.7.3 Overcast and Dense Cloud Conditions . . . 29

5 Conclusion **31**
5.1 Discussion of Findings . . . 31
5.1.1 Theory Changes . . . 31
5.1.2 Process Changes . . . 31
5.1.3 Results Differences . . . 31
5.1.4 Temporal Prediction and Generalization . . . 32
5.1.5 Weather-Dependent Forecasting Fidelity . . . 32
5.2 Summary of Project Progress to Date . . . 32
5.3 Operation Plan . . . 33
5.4 Problems, Obstacles, and Solutions . . . 33
5.5 Conclusion and Future Work . . . 34

List of Figures

2.1 SkyGPT pipeline (source: [1] image by: Yuhao Nie, et al.) 10
2.2 Stochastic video prediction module pipeline (source: [1] image by: Yuhao Nie, et al.) . . 11

4.1 Reconstruction before optimization on SIRTA dataset. 26
4.2 Reconstruction after optimization on SIRTA dataset. 27
4.3 Evolution of PSNR over 40 training epochs for SkyGPT. 28
4.4 SkyGPT odd frame predictions for clear sky conditions from t + 1 to t + 15 28
4.5 SkyGPT odd frame predictions for partly cloud conditions from t + 1 to t + 15 29
4.6 SkyGPT odd frame predictions for dense cloud conditions from t + 1 to t + 15 30

List of Tables

2.1 Performance comparison on the SKIPP'D dataset. 14
2.2 Performance comparison between the diffusion-based model and VAE-based benchmarks. 15

3.1 Dataset properties used for benchmark context. 18
3.2 VQ-VAE hyperparameter comparison between earlier sweeps and the best ajgap configuration. 21
3.3 Fixed SkyGPT network and learning hyperparameters retained for future temporal integration. 22

4.1 Top 5 distinct model configurations ranked by best recorded PSNR from the hyperparameter search summary. 24

Chapter 1

Introduction

1.1 Background and Motivation

Accurate short-term solar irradiance forecasting is important for grid stability, photovoltaic scheduling, and renewable-energy management. In practice, however, sky conditions evolve through highly nonlinear cloud translation, deformation, splitting, and merging. These changes directly affect irradiance at short time scales and make future sky states difficult to predict from raw image sequences alone. A practical forecasting system must therefore preserve both pixel-level reconstruction fidelity and the physical plausibility of evolving cloud structures.

Recent work on SkyGPT [1] addresses this challenge through a hierarchical video-prediction framework that combines discrete latent representation learning with a physics-aware temporal predictor. In that framework, a Vector-Quantized Variational Autoencoder (VQ-VAE) [2] compresses historical sky-image sequences into discrete latent codes, and a downstream temporal model predicts how those codes evolve over time. The quality of the compressed representation is therefore not a peripheral detail: it directly determines whether the later forecasting module can preserve cloud boundaries, motion cues, and scene structure.

This project focuses on the implementation of the SkyGPT pipeline, utilizing it as the primary benchmark. Initial evaluations revealed that several baseline configurations and source code versions suffered from suboptimal reconstruction quality and systemic technical failures. These issues included persistent software bugs and inconsistent outputs that failed to replicate the performance metrics established in the original literature. Subsequent development, however, showed that this weakness was not fundamental to discrete latent modeling concept itself. Instead, performance improved substantially when latent capacity, quantization strategy, downsampling design, and training stability were revised in a more systematic way. In particular, the improved configuration established a much stronger reference point for this study and motivated a more rigorous academic rewrite of the project.

1.2 Problem Statement

The central problem in this study is how to construct a discrete latent representation for cloud-motion sequences that is compact enough for downstream temporal prediction while still preserving the visually and physically meaningful details of sky dynamics. In practical terms, the problem can be divided into two technical challenges.

First, high-dimensional RGB image sequences must be compressed into discrete latent representations without destroying motion-relevant information such as cloud boundaries, texture transitions, and large-scale illumination structure. If the codebook is too small, the hidden representation too narrow, or the quantization process unstable, the reconstructions become overly smooth and lose detail that later prediction modules need.

Second, the learned representation must remain compatible with future prediction in latent space.

A temporally predictive model such as SkyGPT can only be effective when the latent tokens preserve the structure of cloud motion in a stable and interpretable way. Weak reconstruction at the compression stage propagates into all later stages of the forecasting pipeline.

Therefore, this study frames the VQ-VAE not merely as an auxiliary compression block, but as a prerequisite for credible cloud-motion forecasting.

1.3 Objectives and Contributions

The objectives of this project are as follows:

1. Analyze the SKIPP'D and SIRTA sky-image datasets with emphasis on temporal resolution, visual variability, and their implications for discrete latent representation learning.
2. Review the theoretical foundations of generative modeling, variational autoencoders, vector quantization, and physics-informed temporal prediction in the context of cloud-motion forecasting.
3. Impementing VQ-VAE for cloud-image reconstruction and identify the architectural and learning choices that most strongly influence performance.
4. Inference the workflows with full pipeline using VQ-VAE parts and Transformers parts
5. Establish a reproducible pipeline with the temporal prediction of SkyGPT.

1.4 Scope of Work

This study is limited to the following scope:

1. The benchmark environments are the SKIPP'D and SIRTA datasets resized to a common spatial resolution for controlled experimentation.
2. The primary technical focus is the VQ-VAE compression stage of the SkyGPT pipeline rather than full end-to-end video prediction.
3. Model quality is evaluated mainly through validation PSNR and validation loss, supported by qualitative interpretation of reconstruction behavior.
4. Architectural comparison emphasizes codebook size, hidden channels, residual depth, embedding dimension, downsampling strategy, and loss weighting.
5. The downstream SkyGPT temporal module is discussed only to justify why high-quality latent representations are necessary for later prediction.

1.5 Expected Outcomes

The expected outcomes of this project are:

1. A structured benchmark of VQ-VAE performance on the SKIPP'D and SIRTA datasets;
2. A documented reference configuration for the strongest development-branch model;
3. A ranked comparison of leading model configurations with interpretation of their architectural trade-offs; and
4. A technically grounded basis for future integration of the improved latent model into the full SkyGPT stochastic video-prediction pipeline.

Chapter 2

Literature review

2.1 Theoretical Foundations for Generative Compression

This chapter summarizes the theoretical foundation required for cloud-motion benchmarking with the SkyGPT framework. The emphasis is placed on the generative compression stage because the quality of the discrete latent representation directly determines whether downstream temporal prediction can preserve realistic cloud structures.

2.1.1 Overview of the SkyGPT

SkyGPT pipeline

The SkyGPT pipeline transforms historical sky images into probabilistic photovoltaic power forecasts through a two-stage process. Historical sky images captured at 2 minute intervals from **15 minutes before the present time** to **1 minute before the present time** are fed into a stochastic video prediction model, which generates N diverse future scenarios, each containing predicted sky images from **1 minute after the present time** to **15 minutes after the present time**. The final predicted image at **15 minutes after the present time** from each scenario is then passed to a photovoltaic output prediction model that maps the sky conditions to the corresponding power output (photovoltaic output forecasting).

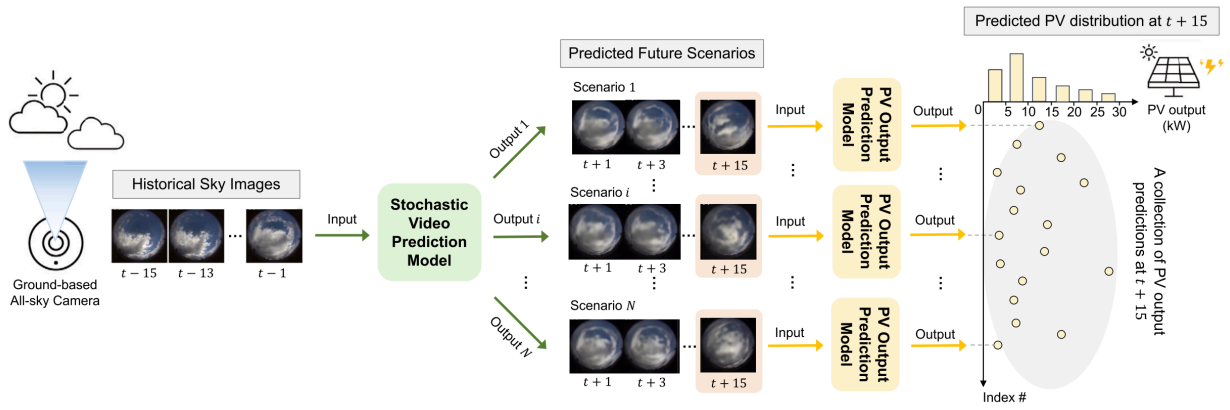

Figure 2.1: SkyGPT pipeline (source: [1] image by: Yuhao Nie, et al.)

Stochastic video prediction module

The SkyGPT stochastic video prediction model is a combination of a Vector Quantized Variational Autoencoder (VQ-VAE) with a transformer constrained by physics (Phy-transformer). The VQ-VAE learns discrete latent representations of video frames, while the Phy-transformer predicts cloud motion dynamics from visual details using a PhyCell module. This architecture enables SkyGPT to capture

realistic cloud motion patterns and generate diverse future video scenarios from the same historical input sequence, accounting for the inherent uncertainty in cloud dynamics.

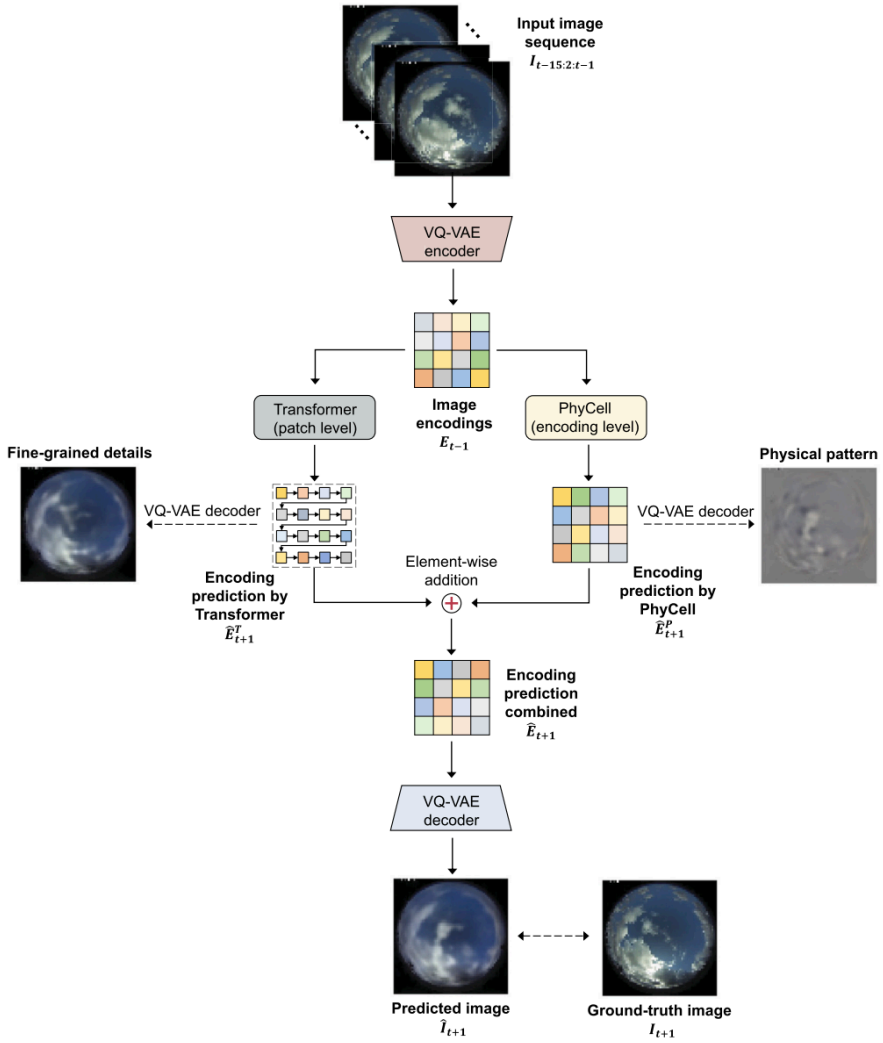

Figure 2.2: Stochastic video prediction module pipeline (source: [1] image by: Yuhao Nie, et al.)

2.1.2 Variational Autoencoders and Quantization

Variational Autoencoders (VAEs) provide a probabilistic framework for learning latent representations by reconstructing observed data through an encoder-decoder architecture [3]. In a conventional VAE, the latent variable is continuous and regularized through a prior distribution. This design is effective for representation learning, but it is less naturally aligned with discrete sequence modeling for autoregressive prediction.

Vector Quantized Variational Autoencoders (VQ-VAE) [2] extend this framework by replacing the continuous latent variable with a finite codebook of learned embeddings. Given an input sequence x, the encoder produces a latent tensor $z_e(x)$, and each latent vector is mapped to its nearest codebook entry $z_q(x)$. The decoder reconstructs the input from the quantized representation. This discrete bottleneck is attractive for cloud-motion modeling because it compresses high-dimensional sky images into symbolic latent tokens that can later be predicted by sequence models.

For cloud imagery, the latent representation must encode both low-frequency illumination structure

and high-frequency cloud boundaries. If the latent capacity is insufficient, the reconstruction becomes overly smooth and loses motion-relevant detail. Conversely, a sufficiently expressive and stable codebook can preserve sharper cloud structures and improve reconstruction quality.

2.1.3 VQ-VAE's loss function

The VQ-VAE framework introduces quantization into the encoder-decoder pipeline while using a straight-through estimator to preserve gradient flow during training. In the present project, the learning objective is best understood as a combination of reconstruction, commitment, and perceptual components:

$$\mathcal{L}_{total} = \beta_{reconstruction}\mathcal{L}_{reconstruction} + \beta_{commitment}\mathcal{L}_{commitment} \tag{2.1}$$

- **The reconstruction component:** encourages pixel-level fidelity. In the improved implementation, this term is based on Charbonnier loss rather than ordinary mean squared error:

$$\mathcal{L}_{Charbonnier} = \sum_i \sqrt{(y_i - \hat{y}_i)^2 + \epsilon^2}, \tag{2.2}$$

where ϵ is a small constant for numerical stability. This choice is useful for cloud imagery because it is less sensitive to outliers since it act like L1 when error is near ϵ and L2 when error is high.

The commitment: component keeps encoder outputs close to their selected codebook vectors:

$$\mathcal{L}_{\text{commitment}} = \|z_e(x) - \text{sg}[z_q(x)]\|_2^2 \tag{2.3}$$

- $z_e(x)$: the output of the encoder (continuous latent space).
- $z_q(x)$: the quantized latent vector (the nearest neighbor in the codebook).
- $\text{sg}[\cdot]$: the stop-gradient operator. This prevents gradients from flowing back into the codebook during this specific calculation, effectively forcing the encoder to "commit" to a representation that the codebook can map easily.

2.2 Temporal Prediction in Discrete Latent Space

2.2.1 Autoregressive Sequence Modeling

After compression into discrete codes, future prediction can be formulated as a sequence-generation problem in latent space. Instead of predicting future RGB frames directly, the model predicts the next sequence of discrete latent tokens conditioned on historical observations. This factorization is appealing because it reduces the dimensionality of the prediction target and lets the temporal model focus on latent dynamics rather than raw pixels.

The SkyGPT architecture employs a Transformer-style attention stack to capture long-range spatial and temporal dependencies [4]. In principle, this design allows the model to represent multiple plausible cloud evolutions, which is important because sky dynamics are inherently stochastic. However, the success of this temporal model depends on the quality of the latent codes it receives from the VQ-VAE. If the encoder discards cloud boundaries or fine motion cues, the temporal predictor cannot recover them later.

2.2.2 Frame Conditioning and Physical Constraints

SkyGPT combines data-driven sequence modeling with physically motivated structure. Historical frames can be encoded through a ResNet34-based conditioning pathway, while the PhyCell module introduces constraints related to physically meaningful motion behavior. This hybrid design attempts to balance flexibility with inductive bias.

The physical branch is commonly described through a moment-based regularization objective:

$$\mathcal{L}_{Phy-transformer} = \mathcal{L}_{moment} + \mathcal{L}_{cross-entropy}, \tag{2.4}$$

where $\mathcal{L}_{moment}$ regularizes the learned filters toward physically meaningful spatial operators and $\mathcal{L}_{cross-entropy}$ measures token-prediction accuracy in latent space. Although full temporal prediction is outside the main scope of the current rewrite, this framework remains important because it explains why strong compression is a prerequisite for end-to-end forecasting.

2.3 Evaluation Criteria

Models are ranked primarily by validation PSNR because the weighted loss values are not directly comparable across all configurations. Validation loss is still reported as a secondary stability indicator, while qualitative reconstruction analysis helps interpret whether higher-scoring models truly preserve cloud boundaries and scene structure. When metrics are close, the benchmark also considers architectural efficiency and reproducibility.

2.4 Related Work

2.4.1 Diffusion Models for High-Resolution Solar Forecasts

Hatanaka et al. (2023) [5] introduced a generative approach for solar irradiance forecasting using score-based diffusion models. Unlike the ultra-short-term focus of SkyGPT[1], which utilizes ground-based sky cameras for 15-minute-ahead predictions, this work addresses the day-ahead horizon through the super-resolution of coarse Numerical Weather Prediction (NWP) data to 0.5 km resolution satellite imagery. The two models exhibit significant architectural and operational differences:

- **Spatial Scale**: SkyGPT operates at a local level for site-specific PV power prediction, while Hatanaka et al. focus on regional irradiance downscaling.
- **Modeling Framework**: SkyGPT utilizes a physics-constrained VQ-VAE and Phy-transformer to model stochastic cloud motion. In contrast, the 2023 study employs score-based diffusion to capture the high-dimensional joint distribution of atmospheric variables, enabling more robust probabilistic ensembles.
- **Forecasting Lead Time**: SkyGPT targets the ultra-short-term (minutes), whereas Hatanaka et al. provide day-ahead (hours to days) forecasts.

While SkyGPT demonstrated a CRPS improvement of 13.2% over baseline CNNs on the SKIPP'D dataset, Hatanaka et al. highlighted the computational efficiency and superior resolution of diffusion models compared to conventional ensemble forecast systems like the Global Ensemble Forecast System (GEFS).

2.4.2 Multimodal for Probabilistic Solar Forecasting

Xiong et al. (2025) [6] introduced a multimodal framework utilizing a Conv-GRU, Neural SDE, and Conditional Diffusion Model (CDM) for sky video prediction. While both models decouple cloud motion from power inference for 15-minute-ahead forecasts, key technical differences distinguish them:

- **Multimodal Input**: Unlike SkyGPT[1], which excludes historical power data to avoid signal overfitting , the Multimodal model integrates historical PV data to resolve mapping ambiguities where similar cloud cover produces different power outputs.
- **Diffusion Architecture**: The Multimodal model replaces SkyGPT's VQ-VAE and physics-informed Transformer with a CDM to generate clearer, more accurate future frames.
- **Comprehensive Uncertainty**: While SkyGPT focuses on aleatoric cloud motion , the Multimodal framework employs Gaussian Process Approximation (GPA) to additionally account for epistemic model uncertainty.

To clarify the interpretation of the reported probabilistic forecasting metrics:

- **Continuous Ranked Probability Score (CRPS):** measures the overall accuracy of a probabilistic forecast by comparing the predicted cumulative distribution against the observed value. A lower CRPS indicates that the predictive distribution is both sharper and better calibrated.
- **Winkler Score (WS):** evaluates the quality of prediction intervals by balancing interval width and coverage. A lower WS indicates a narrower interval with appropriate coverage of the true observation, while large penalties are incurred when the observed value falls outside the predicted interval.

In a comparative analysis using the open-source SKIPP'D dataset, the 2025 multimodal framework demonstrated superior probabilistic forecasting performance over SkyGPT, as reflected by lower CRPS and Winkler Score values:

Metric	SkyGPT-UNet (2024)	Multimodal (2025)	Improvement
CRPS (kW)	2.81	2.44	13.2%
Winkler Score (WS)	26.72	19.62	26.6%

Table 2.1: Performance comparison on the SKIPP'D dataset.

These results indicate that the proposed framework produced predictive distributions that were more accurate and more reliable than those of SkyGPT. Furthermore, the studies confirmed that the diffusion-based video module alone improved the Continuous Ranked Probability Score (CRPS) by approximately 11.1% to 16.4% when replaced with the SkyGPT video prediction module.

2.4.3 Diffusion-Based Multi-Resolution Solar Power Prediction

Rastgoo et al. (2025) [7] introduced a diffusion-based framework for ultra-short-term solar power forecasting, centered on the Cross Branch Visual Informer (CB-ViInf). This architecture utilizes multi-resolution image patches to capture both fine-grained cloud textures and global spatial contexts. Integrates a dual-block Temporal Encoder—employing spatial-temporal ridgelet transform (STRT) and

Multi-Frame Adaptive Singular Value Decomposition (MF-ASVD)—for dynamic noise reduction and feature refinement. To quantify uncertainty, the model applies a custom loss function based on the Evidence Lower Bound (ELBO). While sharing SkyGPT's core philosophy of using generative AI and ground-based sky images. However, this work introduces significant technical deviations:

- **Training pipeline**: Unlike SkyGPT's[1] use of a physics-constrained VAE and VideoGPT, Rastgoo et al. utilize a diffusion-based backbone to mitigate by generating the image sequences.
- **Multi-Resolution Feature Extraction**: While SkyGPT operates on single-scale images, CB-ViInf employs three distinct patch streams (tiny, medium, and big) with cross-attention to capture multi-scale interactions in cloud dynamics.
- **Advanced Temporal Denoising**: The model incorporates signal processing algorithms (STRT and MF-ASVD) to adaptively filter noise in the spatiotemporal domain, which enhances robustness against sudden brightness changes or camera noise.
- **Attention Mechanism**: A hybrid spatio-temporal attention block is used to balance local short-term variations with broader global trends, improving adaptability to rapid cloud movements.

Extensive evaluations demonstrate that this diffusion-based approach provides superior accuracy and calibration compared to VAE-based generative baselines similar to SkyGPT:

Metric	VAE-based Baseline	Proposed Diffusion	Improvement
MAPE (Out-of-Sample)	15.6%	4.9%	68.59%
MSE (Out-of-Sample)	0.031	0.012	61.29%

Table 2.2: Performance comparison between the diffusion-based model and VAE-based benchmarks.

The results indicate that the integration of multi-resolution visual semantics and a diffusion-based probabilistic backbone significantly enhances the reliability of solar power predictions under highly variable sky conditions.

2.4.4 Video Representation Learning

Prior work on video representation learning has explored continuous latent models, discrete latent models, and hybrid approaches. In compression-oriented generative modeling, a key design choice is the balance between reconstruction accuracy and representation efficiency. Loss formulation also matters substantially. Robust reconstruction objectives such as Charbonnier loss are often more effective than plain MSE when edges and transition regions are visually important.

2.4.5 Discrete Representation Learning for Domain-Specific Data

VQ-VAE [1] and related discrete latent models have been applied successfully in speech, vision, and control. However, domain-specific applications require careful adaptation. Cloud imagery differs from natural-image benchmarks because it contains large coherent regions, soft illumination gradients, and intermittently sharp cloud boundaries. As a result, architectural settings and loss balances that work for generic images do not necessarily transfer directly to sky-image reconstruction.

2.5 Evaluation Metrics

The reconstructed or predicted video frames can be evaluated using standard image-quality metrics such as MSE, MAE, and perceptual similarity. In this study, validation PSNR and validation loss are the principal indicators for comparing VQ-VAE configurations.

$$\text{MSE} = \frac{1}{N}\sum_{k=1}^{N}\sum_{i}\left(\mathcal{I}_i^k - \hat{\mathcal{I}}_i^k\right)^2 \tag{2.5}$$

$$\text{MAE} = \frac{1}{N}\sum_{k=1}^{N}\sum_{i}\left|\mathcal{I}_i^k - \hat{\mathcal{I}}_i^k\right| \tag{2.6}$$

$$\text{PSNR} = 10\log_{10}\left(\frac{MAX^2}{\text{MSE}}\right) \tag{2.7}$$

Higher PSNR indicates better reconstruction quality. In addition to these image-level metrics, codebook perplexity and active-code counts are useful for interpreting how efficiently the discrete latent space is being utilized during training.

Chapter 3

Methodology

3.1 Research Workflow

The methodology of this project follows a benchmark-oriented workflow designed to isolate the role of discrete latent representation learning in cloud-motion modeling. The work proceeds through four main stages:

- **Study the SkyGPT framework** and identify the role of generative compression in the overall stochastic video-prediction pipeline.
- **Analyze the SKIPP'D and SIRTA datasets** to understand their temporal and visual characteristics and how those characteristics affect latent representation learning.
- **Benchmark VQ-VAE configurations** across multiple architectural and loss-weighting settings.
- **Consolidate the strongest development-branch implementation** and reinterpret the project around a validated reference configuration rather than an unresolved baseline failure.

Training protocol recommendations To ensure reproducibility and fair comparison across configurations, the training process should explicitly record:

- maximum number of epochs, early-stopping patience, and the validation metric used for model selection;
- optimizer and learning-rate schedule settings;
- mixed-precision and gradient-stability settings when used;
- codebook update strategy, including EMA parameters and commitment warm-up;
- the full configuration object saved alongside each checkpoint and validation log.

3.2 Dataset Description

The SKIPP'D and SIRTA datasets provide sky-image sequences for evaluating short-term cloud-motion modeling under different environmental conditions. Although both datasets are resized to a common spatial resolution for experimentation, they differ in acquisition interval, scene diversity, and atmospheric characteristics. These differences motivate dataset-aware benchmarking rather than assuming that one VQ-VAE configuration will generalize uniformly across both domains.

- SKIPP'D dataset download: `https://purl.stanford.edu/dj417rh1007`
- SIRTA dataset access: `https://sirta.ipsl.fr/data-request/`

Dataset	SKIPP'D	SIRTA
Resolution	64 × 64	64 × 64
Capture interval	1 minute	1–2 minutes
Start date	09/03/2017	01/01/2023
End date	26/10/2019	31/12/2023
Start time	06:00	05:00
End time	Not over 20:00	Not over 22:00

Table 3.1: Dataset properties used for benchmark context.

Beyond these acquisition properties, the model pipeline assumes 16-frame input sequences at a spatial resolution of 64×64 pixels, with RGB channels normalized to a symmetric range around zero before training. The SIRTA configuration further organizes the data through train/validation/test splits and frame lists designed for sequence prediction benchmarking.

3.3 Experimental Design

3.3.1 Modified loss function in VQ-VAE

The reconstruction term is used to preserve pixel-level fidelity between the target image and its reconstruction. In the improved implementation, this component is formulated using the Charbonnier loss instead of the conventional mean squared error:

$$\mathcal{L}_{\text{Charbonnier}} = \sum_i \sqrt{(y_i - \hat{y}_i)^2 + \epsilon^2}, \tag{3.1}$$

where ϵ is a small positive constant introduced for numerical stability. The Charbonnier loss can be interpreted as a differentiable approximation to the ℓ_1 loss. In particular, it behaves approximately quadratically for small reconstruction errors and approximately linearly for large errors. This property makes it more robust to outliers than mean squared error, which is especially desirable in cloud imagery, where localized high-intensity deviations may otherwise dominate the optimization process.

To further improve reconstruction quality, we incorporate a perceptual loss term defined as [8]

$$\mathcal{L}_{\text{perceptual}} = \|\phi(x) - \phi(\hat{x})\|_2^2, \tag{3.2}$$

where $\phi(\cdot)$ denotes a feature extractor that maps an image into a learned representation space, typically using intermediate activations from a pretrained network. Unlike purely pixel-wise objectives, perceptual loss measures discrepancies in higher-level feature representations between the input x and the reconstruction $\hat{x}$. As a result, it encourages reconstructions that better preserve perceptually meaningful structures and visual coherence.

3.3.2 Codebook Stability, Perplexity, and Latent Capacity

An important diagnostic for vector-quantized models is the extent to which the codebook is effectively utilized. Let K denote the number of codebook entries, and let p_k represent the empirical usage

probability of the k-th code:

$$p_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[z_i = k], \qquad \text{with} \qquad \sum_{k=1}^{K} p_k = 1, \tag{3.3}$$

where z_i is the discrete encoding index assigned to the i-th latent position.

Perplexity Codebook utilization is commonly quantified by the perplexity, defined as the exponential of the empirical entropy:

$$\text{Perplexity} = \exp\left(-\sum_{k=1}^{K} p_k \log(p_k + \epsilon)\right), \tag{3.4}$$

where ϵ is a small constant introduced for numerical stability. A higher perplexity indicates a more uniform distribution of code usage across the codebook, whereas a low perplexity suggests that only a limited subset of codes is being used.

Active Codes A complementary diagnostic is the number of active codes, defined as the number of distinct code indices observed in a batch:

$$\text{ActiveCodes} = \left|\{z_i\}_{i=1}^{N}\right| = \left|\text{unique}(z)\right|. \tag{3.5}$$

In implementation terms, this is computed as

$$\text{ActiveCodes} = \text{len}\big(\text{unique}(\text{encodings})\big). \tag{3.6}$$

A well-utilized codebook typically satisfies

$$\text{Perplexity} \approx K \qquad \text{and} \qquad \text{ActiveCodes} \approx K, \tag{3.7}$$

indicating that the latent representation is making effective use of the available discrete capacity. In contrast, low values of these metrics are indicative of codebook collapse and reduced representational diversity. Importantly, increasing K alone does not guarantee improved reconstruction performance, since the encoder and decoder must also possess sufficient capacity to exploit the larger discrete vocabulary.

Interpolated 3D Convolution for Upsampling We define an interpolated 3D convolution operator that decouples upsampling from convolution. Given an input tensor $x \in \mathbb{R}^{B\times C\times T\times H\times W}$ and stride $\mathbf{s} = (s_t, s_h, s_w)$, the temporal dimension is first upsampled by repetition:

$$x'_{b,c,t,:,:} = x_{b,c,\lfloor t/s_t \rfloor,:,:}, \qquad t = 0, \ldots, s_t T - 1. \tag{3.8}$$

The resulting tensor is then spatially upsampled via interpolation:

$$\tilde{x} = \text{Interp}_{(s_h,s_w)}(x'), \tag{3.9}$$

where Interp denotes bilinear interpolation applied independently to each temporal slice [9]. Finally, a 3D convolution with unit stride is applied:

$$y = \text{Conv3D}(\tilde{x}; \theta). \tag{3.10}$$

This design replaces strided transposed convolution with explicit upsampling in both temporal and spatial dimensions prior to convolution. Consistent with prior observations on resizing before convolutional processing, this separation can improve smoothness and reduce undesirable interpolation artifacts [10]. As a result, it improves reconstruction smoothness and helps mitigate checkerboard artifacts.

In the development branch, codebook dynamics are further stabilized through exponential moving-average (EMA) updates together with a warm-up schedule for the commitment loss term. This strategy reduces instability in code assignment and allows the model to explore the latent space more effectively before stronger commitment is enforced.

3.3.3 Hyperparameter Search Strategy

The benchmark draws on two complementary evidence sources:

1. a hyperparameter search summary covering distinct VQ-VAE configurations across datasets; and
2. the final validation record of the best development-branch model, `vqvae_ajgap`.

The search dimensions include codebook size, residual depth, hidden width, embedding dimension, downsampling pattern, learning rate, and the relative weighting of reconstruction, commitment, and perceptual losses. This design makes it possible to compare not only isolated training runs but also the broader interaction among architectural capacity and optimization strategy.

3.3.4 Full Pipeline Integration

Once the optimal hyperparameters for the VQ-VAE model are identified based on the best PSNR results, the model is integrated into the complete SkyGPT pipeline. This integration is performed in two primary steps:

1. **Input Alignment:** The output from the VQ-VAE encoder is matched to the input requirements of the Attention Stack and the PhyCell module. This ensures that the frame conditioning data flows correctly through the temporal and physical prediction layers.
2. **Dimensional Consistency:** To prevent shape errors within the codebase, tensor dimensions are standardized. This involves using permutation (specifically rearranging the data to a (B, C, T, H, W) format) to ensure that the latent representations are compatible across all modules in the pipeline.

3.4 VQ-VAE Architecture and Configuration

3.4.1 Encoder-Decoder Structure

The VQ-VAE used in this project applies 3D convolutional encoding, vector quantization through a discrete codebook, and convolutional decoding back into image-space sequences. The architecture is

designed to compress a short video sequence into a latent representation that remains expressive enough for later temporal modeling.

Earlier sweeps explored compact and medium-scale VQ-VAE settings, mainly with downsampling $(4, 4, 4)$ and hidden widths up to 240 channels. The improved `vqvae_ajgap` configuration adopts a stronger architecture and serves as the main reference model for the rewritten study.

Parameter	**Earlier sweep range**	**Best configuration**
Codebook size (n_codes)	64, 128, 256, 512, 2048	2048
Residual layers (n_res_layers)	1, 4	6
Hidden channels ($n\_hiddens$)	30, 60, 90, 240	384
Embedding dimension	30–256	394
Downsample	$(4, 4, 4)$	$(2, 4, 4)$
Learning rate	10^{-6} to 10^{-1}	1×10^{-4}
Reconstruction weight ($\beta_{reconstruction}$)	1.0, 16.667	1.0
Perceptual weight ($\beta_{perceptual}$)	0.1–1	0.3
Commitment weight ($\beta_{commitment}$)	0.1–1.0	0.2
EMA decay	0.9–0.95	0.95
Restart threshold	0.9–1	1.0

Table 3.2: VQ-VAE hyperparameter comparison between earlier sweeps and the best ajgap configuration.

The improved model increases latent capacity in four ways: a large codebook, deeper residual processing, wider hidden channels, and a larger embedding dimension. In addition, the asymmetric downsampling $(2, 4, 4)$ reduces temporal compression pressure relative to the earlier $(4, 4, 4)$ setting. This is beneficial for cloud-motion sequences because temporal continuity is preserved more faithfully in the latent space.

3.4.2 Loss Function and Optimization Strategy

The improved implementation combines several training design choices that were not as clearly consolidated in earlier experiments:

- Charbonnier reconstruction loss in place of plain MSE;
- a commitment term tied to discrete codebook learning;
- an optional perceptual term to preserve visually meaningful structure;
- EMA-based codebook updates for stable embedding adaptation;
- commitment warm-up in the first training epochs; and
- mixed-precision training with gradient clipping.

From the implementation files, the VQ-VAE training loop uses Adam optimization, automatic mixed precision, gradient clipping at max norm 1.0, and a cosine annealing learning-rate schedule. The commitment term is warmed up over the first 30 epochs before reaching its target value. These details are methodologically important because the final benchmark is not explained by model size alone; it also depends on more stable optimization and better-balanced loss design.

3.5 SkyGPT Temporal Prediction Context

Although the main focus of this study is the compression stage, the benchmark is still situated within the broader SkyGPT framework. The downstream temporal module includes a Transformer attention stack, positional and frame conditioning, and a PhyCell-inspired physical constraint pathway. The fixed context used for future integration is summarized below.

Category	Parameter	Initial value
Network architecture	Normalization std (normal_std)	0.02
	Kernel size (kernel_size)	7
	ResNet34 downsample (resnet34_sample)	$(1, 4, 4)$
	Attention dropout (attn_dropout)	0.3
	Hidden dropout (dropout)	0.2
	Attention heads (heads)	4
	Hidden dimension (hidden_dim)	576
	Layers (layers)	8
Input configuration	Precision (precision)	32
	Conditional frames (n_cond_frames)	16
Learning	Learning rate (learning_rate)	1×10^{-3}
	Learning betas (learning_betas)	$(0.9, 0.999)$

Table 3.3: Fixed SkyGPT network and learning hyperparameters retained for future temporal integration.

The present study does not claim end-to-end superiority of the whole SkyGPT system. Instead, it establishes that the improved VQ-VAE in the development branch is sufficiently strong to serve as a credible latent representation module for subsequent cloud-motion prediction experiments.

3.6 New Sky Image Dataset at CUEE

To manage the high-volume video data captured from the CUEE rooftop, we developed a specialized codebase focused on data extraction and recording efficiency. This pipeline converts high-bitrate raw video into synchronized, resized image sets, effectively reducing storage requirements while maintaining data integrity. The process follows a systematic three-stage algorithm designed to minimize file sizes and optimize the dataset for machine learning workflows:

- **Temporal Downsampling:** To eliminate temporal redundancy, the system implements a condi-

tional sampling logic. Using OpenCV and datetime metadata, the algorithm only extracts frames that align with a specific interval (defined via YAML configuration) and a strict synchronization constraint: frames must occur at an even-minute mark and the zero-second point. This ensures the resulting dataset is temporally consistent across long-term observation periods.

- **Spatial Normalization:** Every frame is resized to a standard 1080×1080 square. This reduces the bitrate of the original MP4 stream and standardizes the input dimensions for subsequent analysis, significantly lowering the memory footprint of each individual image.
- **Serialized Batch Processing:** The pipeline is designed for idempotent batch processing, automatically scanning input directories and skipping previously processed files to ensure efficiency. Extracted frames are stored as NumPy arrays within Pandas Pickle (.pickle) files. This binary serialization avoids the overhead of thousands of individual image files and allows for high-speed data loading during the modeling phase.

By converting high-bitrate video into synchronized and resized image sets, this method successfully reduced the raw data from **400–500 GB** to only **10–20 GB**. This is a reduction of about 95% without sacrificing the essential visual features.

Chapter 4

Results and discussion

4.1 Overview of Benchmark Evidence

This chapter summarizes the latest reconstruction benchmark evidence from two sources: the aggregated hyperparameter tunings and the final validation record of the best development-branch model. The discussion emphasizes distinct model configurations rather than top individual epochs, because configuration-level comparison is more meaningful for academic benchmarking.

4.2 Top Performing Model Configurations

Rank	PSNR	Loss	codes	RL	hid	emb	β_{rec}	β_{com}
1	37.2580	0.013486	2048	6	384	394	1	0.20
2	37.0842	0.006936	2048	4	240	256	16.667	0.30
3	36.5880	0.030966	2048	4	240	256	4.000	0.30
4	35.9608	0.016386	2048	6	192	192	1.000	0.20
5	35.9160	0.009440	2048	4	240	256	16.667	0.30

Table 4.1: Top 5 distinct model configurations ranked by best recorded PSNR from the hyperparameter search summary.

Interpretation The top-ranked configurations reveal a clear capacity trend. Models with large codebooks and wider latent channels generally achieve higher PSNR, indicating that cloud reconstruction benefits from richer discrete representations. At the same time, the comparison shows that raw loss values cannot be interpreted independently of their weighting scheme: models with lower weighted loss do not necessarily achieve the best PSNR. For that reason, PSNR is treated as the main ranking metric in this study. The final configuration improved through a more expressive and better-stabilized architecture which use maximum capabilities of the hardware setting.

4.3 Sensitivity Analysis

4.3.1 Architectural Capacity

A consistent pattern across the benchmark is that larger codebooks and deeper latent-processing networks improve reconstruction quality. Small-codebook models can still be competitive when efficiency is important, but they generally saturate at lower PSNR. In contrast, models with 2048 codes and stronger hidden capacity preserve sharper boundaries and more detailed cloud structure.

This observation supports the interpretation that cloud-motion reconstruction is not a trivial compression task. The latent representation must preserve gradual illumination structure, broad cloud

masses, and fine transition zones simultaneously. These demands are easier to meet when the model has sufficient discrete vocabulary and processing depth which require large computaional cost.

4.3.2 Loss Balancing and Optimization

The benchmark also shows that optimization strategy matters as much as raw model size. A large model without stable quantization dynamics may still underperform. The stronger results in the development branch are associated with several mutually reinforcing choices:

- Charbonnier reconstruction loss instead of plain MSE;
- EMA-based codebook updates;
- a commitment-loss warm-up schedule;
- mixed-precision training with gradient clipping
- an added perceptual component to support visually meaningful reconstruction.
- interponation convolution instead of plain convolution to patch hardware limitation

These improvements explain why the final model can outperform original SKYGPT's VQVAE implementation regard of hardware limitation and robustness.

4.3.3 Dataset Differences

The SKIPP'D and SIRTA datasets exhibit notably different responses during training and inference. Empirical observations suggest that SIRTA contains greater variability in cloud structures and sky conditions, providing richer patterns for the model to learn. In contrast, SKIPP'D includes a higher proportion of clear-sky images, which reduces the diversity of cloud features and makes it more challenging for the VQ-VAE to effectively learn meaningful cloud representations.

As a result, later-stage tuning is primarily conducted on SIRTA, where the model benefits from the increased complexity and variability of the data. This discrepancy highlights the importance of dataset-aware analysis, as hyperparameter configurations that perform well on one dataset may not generalize effectively to another.

4.4 Qualitative Interpretation of the VQ-VAE Problem

The earlier draft treated VQ-VAE as an unresolved failure because many baseline configurations plateaued early and produced visually weak reconstructions. The updated evidence changes that conclusion. The development run continues improving into the later epochs and reaches 37.3216 dB validation PSNR while also achieving a best validation loss of 0.0134865. Therefore, the primary issue is better characterized as a configuration and implementation limitation, rather than an inherent weakness of discrete latent modeling, as was implicitly suspect in the original SKYGPT implementation..

The revised implementation addresses the earlier weakness in three main ways. First, the latent space is made more expressive through a larger codebook, deeper residual processing, and wider hidden channels. Second, the quantization process is stabilized through EMA codebook updates and a better-balanced commitment term. Third, the objective function incorporates a perceptual term, helping the decoder preserve meaningful cloud boundaries instead of optimizing only for smooth pixel averages.

4.5 Figures for Benchmark Review

The following figures remain valuable as visual support for the rewritten chapter and should be retained in the final report where available:

- combined validation PSNR and loss comparisons for SIRTA and SKIPP'D;
- sample reconstructions from the best-loss and best-PSNR checkpoints; and
- correlation plots for SKIPP'D16x15 and SIRTA16x15.

These figures should not be interpreted as standalone evidence of end-to-end forecasting superiority. Rather, they serve as qualitative and diagnostic support, showing that improved quantization design and better-tuned architectures can materially enhance reconstruction quality for complex, non-linear cloud motion.

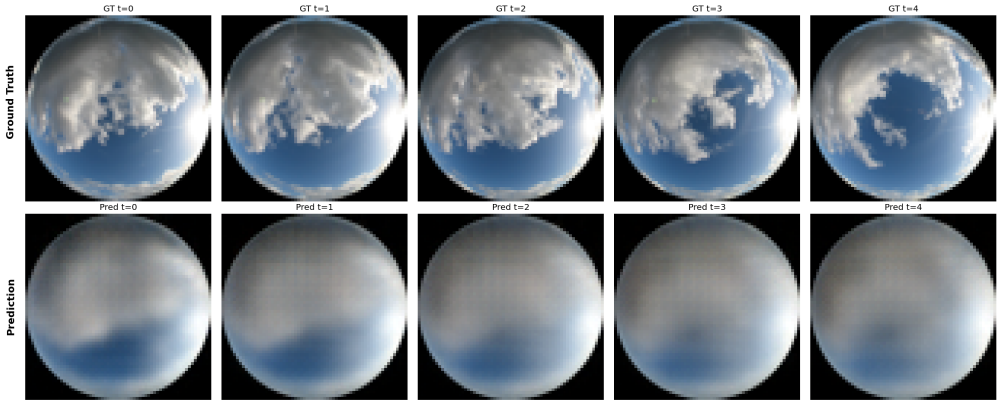

Figure 4.1: Reconstruction before optimization on SIRTA dataset.

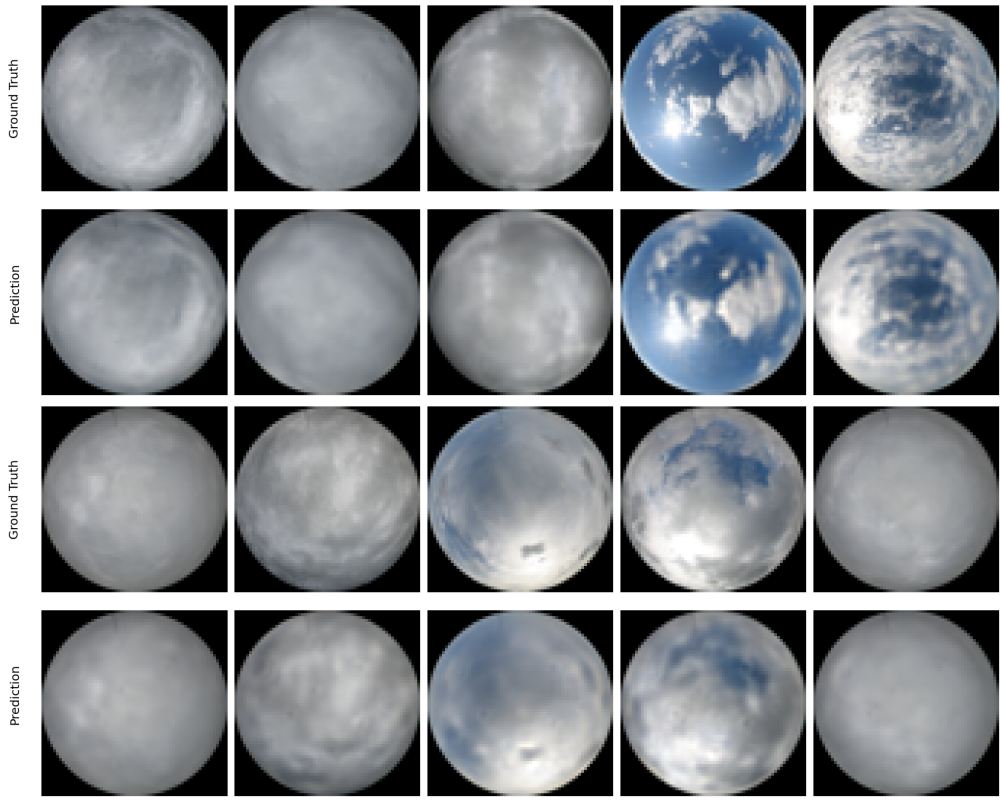

Figure 4.2: Reconstruction after optimization on SIRTA dataset.

4.6 SkyGPT Training Dynamics and Reconstruction Performance

- **Training PSNR:** The training Peak Signal-to-Noise Ratio (PSNR) demonstrates a consistent upward trajectory, featuring a distinct performance shift at Epoch 22 and ultimately converging at a maximum value of 30.73 dB.
- **Validation PSNR:** The validation PSNR initially exhibits high volatility, reflecting the stochastic nature of modeling complex cloud formations, before stabilizing at 30.53 dB after Epoch 30. Notably, the results become steady after this point, and the model achieves its peak predictive performance within this stabilized stage at Epoch 36.

This resulting generalization gap of ≈ 0.20 dB indicates that the model balances high-fidelity reconstruction with the requisite diversity for plausible 15-minute-ahead sky image forecasting. While PSNR serves as a primary metric for reconstruction accuracy derived from Mean Squared Error (MSE).

The graph omits Epochs 1–9 because these initial training steps produced high-outlier PSNR values that drastically skew the visual scale. These artificially high initial metrics are a known mathematical artifact of Mean Squared Error (MSE)-based loss; during early training, the model generates highly blurry, low-frequency images. While these featureless, gray outputs achieve low pixel-wise error compared to clear sky or average cloud cover, they are semantically meaningless and lack the high-frequency detail required for functional forecasting.

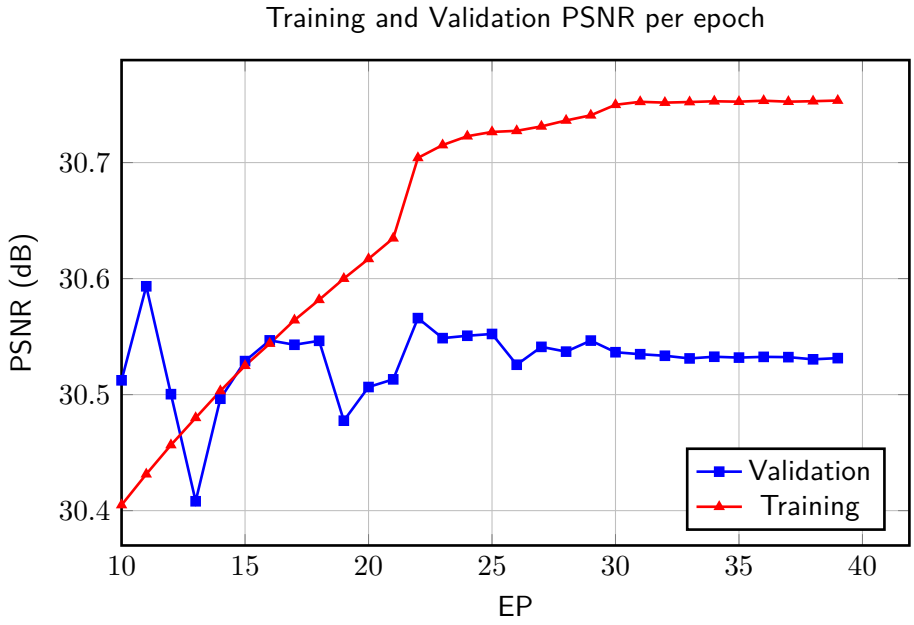

Figure 4.3: Evolution of PSNR over 40 training epochs for SkyGPT.

4.7 SkyGPT Reconstruction Performance

4.7.1 Clear Sky Conditions

In scenarios characterized by a total absence of cloud cover, SkyGPT demonstrates high structural accuracy and consistency, establishing a robust baseline for solar irradiance forecasting:

- **High-Precision Solar Trajectory Tracking:** The model accurately captures the sun's movement across the celestial hemisphere. It maintains high geometric precision regarding solar zenith and azimuth angles, ensuring the primary irradiance source is correctly positioned throughout the $t+15$ forecast horizon without spatial drifting.
- **Radiometric Stability and Noise Mitigation:** Unlike standard generative models that may introduce flickering in static backgrounds, SkyGPT maintains a smooth, temporally consistent sky gradient. This stability indicates that the stochastic components of the framework correctly identify the lack of cloud motion.

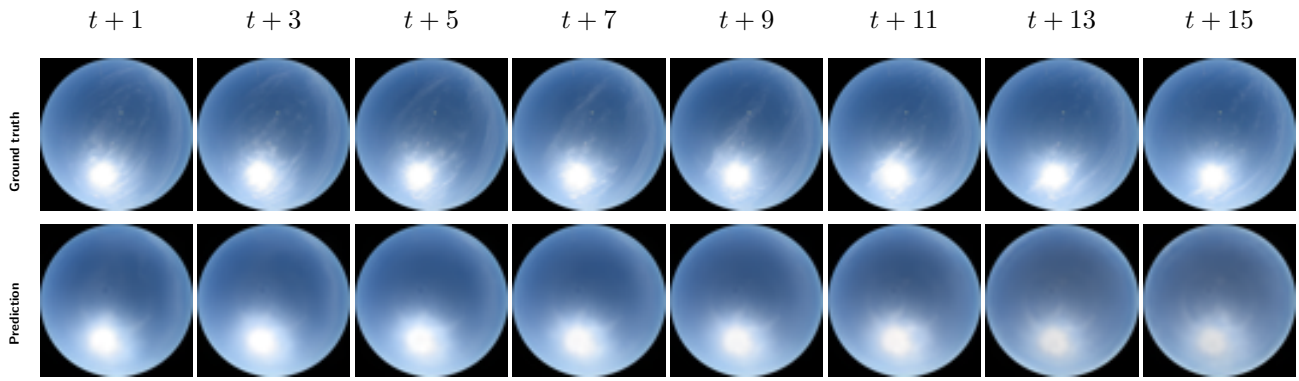

Figure 4.4: SkyGPT odd frame predictions for clear sky conditions from t + 1 to t + 15

4.7.2 Partly Cloudy and Dynamic Cloud Motion

In scenarios involving discrete cloud formations against a clear sky background, the model exhibits its highest predictive fidelity:

- **Precise Dynamic Capture:** SkyGPT effectively captures complex cloud dynamics, including the continuous deformation, condensation, and evaporation of discrete clouds as they move across the field of view.
- **Superior Visual Sharpness:** Unlike deterministic baselines that produce blurry "average" images to minimize pixel-wise loss, the stochastic nature of SkyGPT allows it to generate sharper, high-fidelity images with realistic fine-grained details.
- **Accurate Motion Trends:** The model excels in extrapolating the general trend of cloud movement while simultaneously accounting for inherent stochasticity, maintaining a high degree of perceptual alignment with the ground truth.
- **Enhanced Shading and Contrast:** In high-contrast "clear sky" scenarios, SkyGPT demonstrates improved modeling of light and shading, which provides the downstream PV output predictor with more accurate features for power generation mapping.

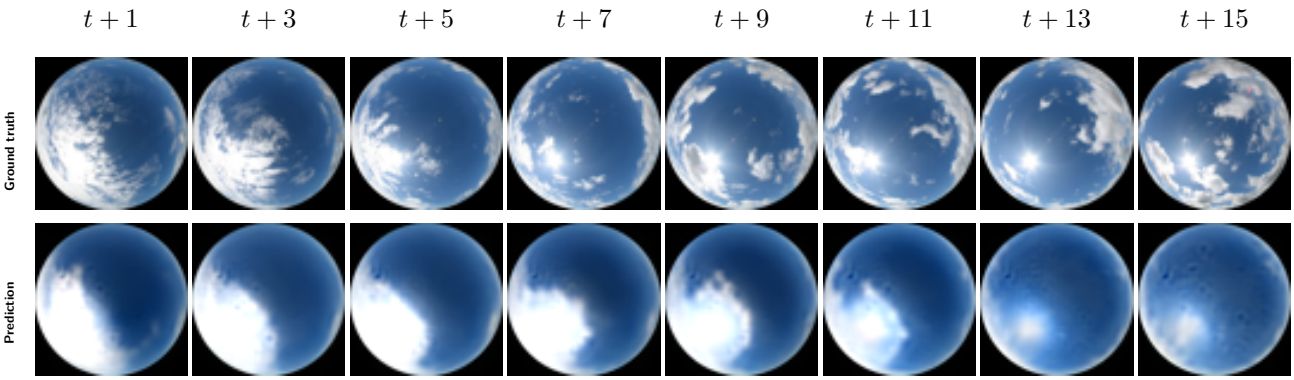

Figure 4.5: SkyGPT odd frame predictions for partly cloud conditions from t + 1 to t + 15

4.7.3 Overcast and Dense Cloud Conditions

Under conditions of dense or thick cloud cover, the model's performance exhibits certain visual limitations:

- **Reduced Visual Fidelity:** The model demonstrates lower performance in maintaining high-frequency spatial details under overcast conditions. This results in predicted images that appear significantly blurrier and more homogeneous compared to the textured ground truth.
- **Averaging of Stochastic Textures:** Because the model struggles to pinpoint specific cloud dynamics in low-contrast environments, it tends to average out plausible future states. This leads to a "smeared" or blurry visual output that captures the general light level but fails to reproduce the complex structural details of the cloud ceiling.

- **Impact of Pixel-wise Loss:** The persistence of blurriness in these scenarios suggests that even with a stochastic approach, the underlying pixel-wise training objectives may still encourage the generation of safe, averaged predictions in high-uncertainty, dense cloud scenarios.

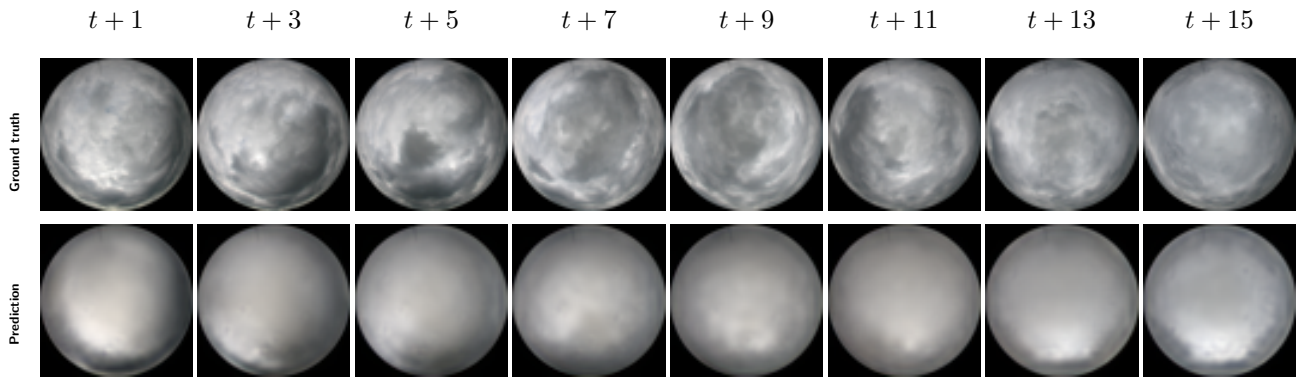

Figure 4.6: SkyGPT odd frame predictions for dense cloud conditions from t + 1 to t + 15

Chapter 5

Conclusion

5.1 Discussion of Findings

The rewritten benchmark supports a clearer interpretation of the project than the earlier draft. The main conclusion is not that VQ-VAE is unsuitable for cloud-motion reconstruction, but that earlier baseline settings did not provide enough latent capacity or sufficient training stability. Once the architecture, downsampling, codebook behavior, and loss design were revised, the model produced substantially stronger reconstructions.

This finding aligns with the main logic of the prototype structure. The VQ-VAE stage is best understood as a generative compression problem whose output directly determines the quality of any later latent-space prediction. In this sense, the strongest contribution of the present work is methodological: it turns a loosely connected set of implementation trials into a coherent benchmark that explains why some configurations fail and why others succeed.

5.1.1 Theory Changes

The theoretical emphasis of the development branch is stronger in two respects. First, it treats VQ-VAE explicitly as a generative compression model whose discrete latent space must preserve motion-relevant cloud structure before temporal forecasting can be trusted. Second, it broadens the learning objective from simple reconstruction and commitment terms to a more balanced formulation that includes perceptual quality and stabilized codebook adaptation. These changes make the proposal more rigorous because they connect model design directly to observable reconstruction behavior.

5.1.2 Process Changes

The development branch improves the experimental process by consolidating updated configuration files, preserving validation logs, and separating sweep-level evidence from the final best-model artifact. In particular, the project now has a clearer benchmark path from the earlier hyperparameter registry to the final `vqvae_ajgap` validation run, allowing the proposal to cite both broad search evidence and the definitive best model.

5.1.3 Results Differences

The main quantitative result difference is that the development branch produces a stronger final VQ-VAE than the earlier benchmark sweep. The best distinct sweep configuration reaches 37.0842 dB PSNR, while the final model reaches 37.3216 dB. In addition, the best validation loss of the final model is 0.0134865. These results show that the development branch does not merely restate the earlier search; it improves the final benchmark through a better architecture and more stable training configuration.

5.1.4 Temporal Prediction and Generalization

The integration of the improved latent model into the SkyGPT temporal stack reveals a stable learning trajectory for 15-minute-ahead forecasting. The training PSNR converged at a maximum of 30.73 dB , while the validation PSNR stabilized at 30.53 dB after Epoch 30. The resulting generalization gap of approximately 0.20 dB indicates that the framework successfully balances high-fidelity reconstruction with the diversity required for plausible stochastic video forecasting. These metrics confirm that the refined VQ-VAE serves as a reliable technical foundation for the downstream Transformer-based predictor.

5.1.5 Weather-Dependent Forecasting Fidelity

Qualitative analysis of the integrated pipeline shows that SkyGPT's predictive performance varies significantly across different sky regimes:

- **Partly Cloudy Conditions:** The model exhibits its highest predictive fidelity in scenarios involving discrete cloud formations. It effectively captures complex dynamics such as cloud deformation, condensation, and evaporation, maintaining high perceptual alignment with the ground truth.
- **Overcast and Dense Clouds:** Under thick cloud cover, the model's performance shows limitations in maintaining high-frequency spatial details. The results appear significantly blurrier as the model tends to average out plausible future states in high-uncertainty environments.
- **Impact of Pixel-wise Loss:** The persistence of blurriness in dense scenarios suggests that underlying pixel-wise training objectives still encourage the generation of "safe," averaged predictions in high-uncertainty conditions.

5.2 Summary of Project Progress to Date

This project has progressed from initial reproduction of the SkyGPT framework to a focused benchmark study of its VQ-VAE component on the SKIPP'D and SIRTA datasets. The earlier phase established the data-processing pipeline, performed hyperparameter search, and identified several limitations in the baseline implementation. The current phase consolidates those findings by adopting the improved `vqvae_ajgap` model from the `dev_CloudMotionBench` branch as the definitive reference configuration.

This transition is academically important because it changes the interpretation of the project. The earlier issue was not that discrete latent modeling was unsuitable for cloud-motion reconstruction, but that the previous implementation and hyperparameter choices were insufficient. The development branch demonstrates that stronger latent capacity, revised downsampling, EMA-based quantization updates, and a perceptual loss term can substantially improve reconstruction quality. As a result, the project now has a validated VQ-VAE benchmark that can support future temporal prediction studies.

5.3 Operation Plan

Work process	2025					2026			
	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
1. Understanding SkyGPT model fundamentals									
2. Data exploration on SIRTA and SKIPP'D datasets									
3. VQ-VAE development with parameter tuning and algorithm optimization									
4. Consolidate best ajgap configuration and verify benchmark evidence									
5. Integrate improved latent model into the stochastic video-prediction workflow									
6. Final analysis and proposal revision based on benchmark results									
7. Write an academic report									

: Current : Plan

5.4 Problems, Obstacles, and Solutions

Problem

- The initial VQ-VAE configurations adapted from the earlier SkyGPT workflow produced limited reconstruction quality and encouraged the incorrect conclusion that the latent model itself was inadequate.
- Many baseline comparisons mixed configuration effects with weighted-loss effects, making it difficult to determine whether poor performance came from architecture, optimization, or metric interpretation.

Proposed Solution

- Replace the unresolved failure narrative with a benchmark grounded in the verified `vqvae_ajgap` implementation from the `dev_CloudMotionBench` branch.
- Use the improved configuration with $n\_codes = 2048$, $n\_res\_layers = 6$, $n\_hiddens = 384$, embedding dimension 394, downsampling $(2, 4, 4)$, learning rate 1×10^{-4}, $\beta_{reconstruction} = 1.0$, $\beta_{perceptual} = 0.3$, and $\beta_{commitment} = 0.2$ as the new technical reference point.

- Report results using the best validation loss and the top five distinct model configurations ranked by PSNR, rather than reporting only the top epochs from a single run.
- Preserve validation logs, configuration files, and checkpoints together so that future stages of the EE409 project remain reproducible and auditable.

5.5 Conclusion and Future Work

The updated benchmark demonstrates that VQ-VAE can serve as an effective representation learner for cloud-motion data when its architecture and quantization process are appropriately designed. The strongest evidence is the `vqvae_ajgap` model, which achieves 37.3216 dB validation PSNR and 0.0134865 validation loss on the SIRTA setting. Beyond the compression stage, the full SkyGPT implementation proves capable of extrapolating general motion trends while maintaining perceptual alignment with ground truth, particularly in partly cloudy conditions.

The primary technical lesson is that non-linear cloud motion requires sufficient latent capacity, stable codebook adaptation, and a loss function that balances numerical reconstruction with perceptual structure preservation. Future work should extend this foundation in three directions:

- First, the optimized VQ-VAE should be further refined to reduce blurring artifacts observed during dense overcast conditions within the SkyGPT temporal stack.
- Second, additional temporal and perceptual evaluation metrics should be introduced to complement PSNR and validation loss.
- Third, broader multi-site experiments should examine how well the learned discrete representations generalize across different climates and seasons.

Bibliography

[1] Y. Nie, E. Zelikman, A. Scott, Q. Paletta, and A. Brandt, "Skygpt: Probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained videogpt," *Advances in Applied Energy*, vol. 14, 7 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666792424000106

[2] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: https://arxiv.org/pdf/1711.00937

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations (ICLR)*, 2014.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] G. G. G. T. P. S. Yusuke Hatanaka, Yannik Glaser, "Diffusion models for high-resolution solar forecasts," *Machine Learning*, 2 2023. [Online]. Available: https://arxiv.org/pdf/2302.00170

[6] X. Z. Z. S. J. F. D. C. D. Z. Binyu Xiong, Yuntian Chen, "Multimodal ultra-short-term probabilistic solar power forecasting with generative ai and transformer," *Advances in Applied Energy*, 12 2025. [Online]. Available: https://doi.org/10.1016/j.adapen.2025.100250

[7] R. S. S. M. M. Razieh Rastgoo, Nima Amjady, "A diffusion-based probabilistic ultra-short-term solar power prediction using the sky image sequences," *IEEE Access*, 12 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11131182

[8] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *arXiv preprint arXiv:1603.08155*, 2016.

[9] D. Han, "Comparison of commonly used image interpolation methods," in *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*. Atlantis Press, Mar. 2013, pp. 1556–1559.

[10] M. Hashemi, "Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation," *Journal of Big Data*, vol. 6, no. 1, p. 98, 2019.