Grounding World Models: Progressive GANs for Physically Plausible Video Generation in Low-Data Regimes

Dhruv Sheth
California Institute of Technology
dsheth@caltech.edu

Abstract

Building effective world models for embodied AI systems requires generating physically plausible temporal sequences, yet current approaches face significant barriers in specialized domains. Diffusion models, while producing high visual quality, demand extensive datasets and computational resources that limit deployment in real-world embodied applications. This paper argues for revisiting Generative Adversarial Networks (GANs) as architecturally-informed alternatives for world modeling in data-constrained environments. We demonstrate this through PhenoGAN, a spatio-temporal Progressively Growing GAN that learns physical process dynamics from limited data ($<$ 1k images). The progressive architecture's inductive bias through its coarse-to-fine learning curriculum mirrors natural developmental processes, enabling physically consistent temporal evolution without massive datasets. We introduce domain-specific evaluation protocols that assess physical plausibility through biological indices, achieving near-perfect correlation ($r > 0.98$) with ground truth measurements across multiple environmental conditions. Our results demonstrate that structurally-informed GANs can achieve both high physical fidelity and computational efficiency, positioning them as complementary tools for world models in embodied AI applications ranging from robotics to automated systems.

1 Introduction

Generating physically plausible dynamic scenes is a cornerstone for building world models for embodied agents Zhu et al. [2024] Fu et al. [2024] Huang et al. [2025a]. While diffusion models have shown success in visually rich simulations Ho et al. [2020], Rombach et al. [2022], Ho et al. [2022] Zhang et al. [2024], their reliance on massive datasets and struggles with maintaining physical consistency Liu et al. [2025], Clark et al. [2019] Zhao et al. [2025] present critical barriers for embodied AI in data-scarce domains Brooks et al. [2024]. This challenge highlights the need for models that can be effectively *grounded* in physical reality, particularly when data is limited Qin et al. [2024].

In this work, we argue that a path toward grounded world models lies in leveraging architectures with strong, task-aligned inductive biases Yin et al. [2025] Aldausari et al. [2020]. We focus on the Progressively Growing GAN (PGGAN) architecture Karras et al. [2018b], whose coarse-to-fine learning curriculum acts as a powerful architectural prior Karras et al. [2018a]. Unlike monolithic models, this structure allows the model to first learn fundamental, low-frequency dynamics before refining high-frequency details Karras et al. [2018b] Sagar [2025]. This process mirrors natural developmental stages, enabling the generation of physically plausible sequences from sparse data Liu et al. [2024] Liu and Vahdat [2025].

Preprint.

We ground this approach with PhenoGAN, a spatio-temporal model that learns plant growth dynamics from fewer than 1k images. Crucially, beyond generation, we introduce a domain-specific evaluation protocol that assesses physical plausibility against biological indices rather than relying solely on visual fidelity. Our results validate that structurally-informed GANs are not only data-efficient but can achieve the high physical fidelity required for grounded world models Skorokhodov et al. [2022], Lin et al. [2025], Karras et al. [2019, 2020b], Huang et al. [2025b], Karras et al. [2020a] Munoz et al. [2020].

2 Methodology

PhenoGAN is a spatio-temporal generative model designed to predict future video frames conditioned on a sequence of past observations. The framework extends the Progressively Growing GAN (PGGAN) architecture Karras et al. [2018b] and is based upon Aigner and Körner [2018], originally conceived for high-quality single image synthesis, to the complex task of video prediction. By learning directly from raw RGB pixel sequences, the model captures the underlying dynamics of physical processes without relying on domain-specific feature engineering. The model consists of a generator and a discriminator trained in an adversarial setting.

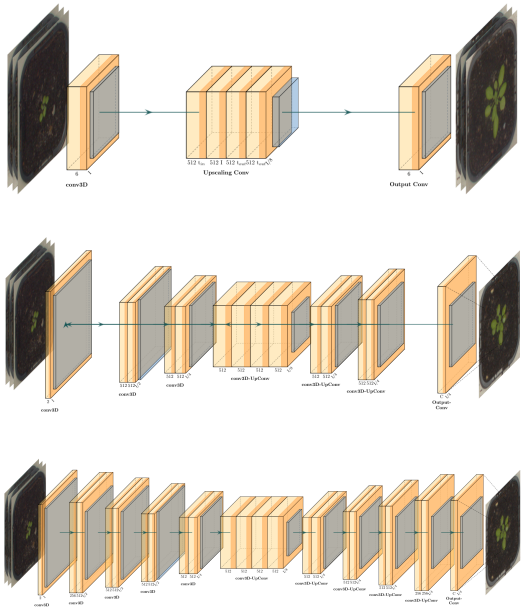

Figure 1: The PhenoGAN generator architecture. New layers are progressively added to increase the output resolution from 4×4 (top) to 128×128 (bottom), forming a coarse-to-fine learning curriculum.

2.1 Architecture and Training

The generator uses an encoder-decoder architecture transforming past frames into future frames. All layers employ spatio-temporal 3D convolutions Tran et al. [2015] to capture appearance and motion. The encoder processes t_{in} frames using 3D convolutions with asymmetric kernels and strides for spatial downsampling, producing compact latent representations. The decoder uses transposed 3D convolutions for upsampling to generate t_{out} future frames, with LeakyReLU activation for vanishing gradient mitigation. The discriminator distinguishes real from fake sequences, providing adversarial signal. For training stability, a mini-batch standard deviation layer computes feature standard deviations across spatio-temporal locations and batch examples, with averaged values replicated and concatenated as additional input features to incentivize varied generator outputs and prevent mode collapse.

PhenoGAN adopts the core PGGAN training strategy, illustrated in Figure 1. Training begins at a low spatial resolution (4×4 pixels), allowing the network to first learn coarse, low-frequency features. As training stabilizes, new layers are added to both generator and discriminator to double the working resolution through smooth transition phases. The full visual progression of generated samples throughout this training process is detailed in Appendix A.5.

PhenoGAN employs the Wasserstein GAN with gradient penalty (WGAN-GP) loss function Gulrajani et al. [2017] to optimize both the generator and discriminator. This loss function is known for improving training stability and the quality of generated samples compared to earlier GAN objectives. The full loss formulation and hyperparameter details are available in the Appendix.

3 Results and Evaluation

We evaluated PhenoGAN on two small plant phenotyping datasets, *Arabidopsis thaliana* and *Beta vulgaris*, to assess physically plausible sequence generation from sparse data. Detailed descriptions of these datasets, including growth and stress conditions, are provided in Appendix B.1. Evaluation combines image fidelity metrics with domain-specific biological indices.

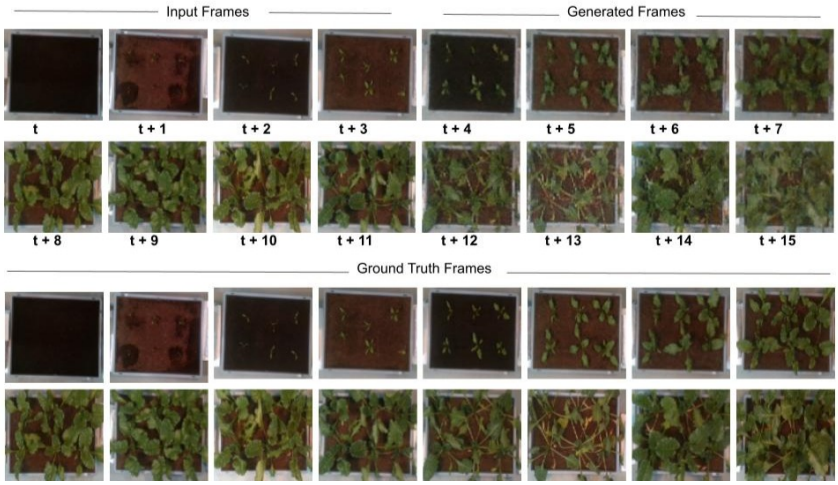

Figure 2: PhenoGAN sequences for *Beta Vulgaris* under combined stress (Drying, Medium Nitrogen, High Weed). **Top:** Input frames (t to $t + 3$). **Bottom:** Generated future frames ($t + 4$ to $t + 7$).

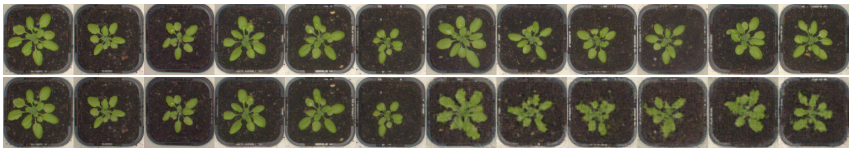

Figure 3: Ground truth vs. PGGAN sequences. **Top:** Ground truth input and expected output. **Bottom Left:** Input to PGGAN. **Bottom Right:** PGGAN output.

3.1 Quantitative Evaluation of Physical Plausibility

PhenoGAN generates visually coherent growth sequences. As shown in Figure 2 and Figure 3, when modeling *Beta vulgaris* under combined stresses (Drying, Medium Nitrogen, High Weed), the model captures complex morphological changes: foliage density increases and plant health shifts.

While visual quality is a prerequisite, we argue that for a world model to be useful for embodied agents, its generations must be physically plausible. To validate this, our evaluation protocol moves beyond standard pixel-level metrics. We first establish a baseline with common image fidelity scores and then introduce a framework for quantifying biological plausibility using domain-specific vegetation indices.

Image Fidelity: As shown in Table 1, PhenoGAN achieves strong performance across standard metrics.

Table 1: **Image Fidelity and Distributional Similarity Metrics for PhenoGAN.** Lower is better for MSE and FID; higher is better for SSIM and PSNR.

Dataset	MSE ↓	SSIM ↑	PSNR ↑	FID ↓
Arabidopsis	0.0098	0.562	26.32	N/A
Vulgaris	0.0077	0.728	27.42	116.73

Biological Plausibility via Vegetation Indices: We evaluated biological understanding using vegetation indices—mathematical combinations of RGB channels that quantify plant health:

Index	Formula	Purpose
ExG	2g - r - b	Highlights green vegetation
ExR	1.4r - g	Indicates non-vegetated/stressed areas
ExGR	ExG - ExR	Robust vegetation segmentation index
PLA	-	Projected Leaf Area from ExGR segmentation

PhenoGAN achieved near-perfect correlations with ground truth across all indices ($r \geq 0.99$ for *Beta vulgaris*), including perfect correlation for ExR ($r = 1.00$). Performance remained high for both control ($r = 0.997$ for PLA) and stressed plants ($r = 0.992$ for PLA), demonstrating quantitatively accurate responses to environmental stressors.

This high performance is consistent across all evaluation subsets. A detailed quantitative analysis, including correlation scatter plots and a performance breakdown for each individual stress condition for the *Beta vulgaris* dataset, is available in Appendix B.4.

Table 2: **Pearson's Correlation Coefficient (r)** between Generated and Ground Truth Vegetation Indices for PhenoGAN, indicating physical plausibility. Results are statistically significant.

Dataset	PLA ↑	ExG ↑	ExGR ↑	ExR ↑
Arabidopsis	0.98	0.97	0.96	0.83
Vulgaris	0.99	0.99	0.99	1.00
- Control Avg.	*0.997*	*1.00*	*0.99*	*0.99*
- Stress Avg.	*0.992*	*0.98*	*0.98*	*0.98*

Visualizations of these indices applied to sample frames from both datasets, demonstrating their effectiveness in segmenting plant biomass, are provided in Appendix B.3.

4 Conclusion

This paper has argued for the renewed importance of data-efficient GANs in building physically-grounded world models for embodied AI. We grounded this position through a detailed case study, PhenoGAN, a spatio-temporal PGGAN that successfully models complex plant growth and stress response dynamics from $<$ 1k images. The success stems from the powerful inductive bias of the progressive growing architecture—its coarse-to-fine learning curriculum mirrors natural developmental processes, enabling plausible dynamics learning from temporally sparse data. The model's physical plausibility is validated by near-perfect correlations ($r > 0.98$) with ground-truth biological markers and superior performance over feature-based state-of-the-art methods. While the architectural prior is highly effective for developmental processes, its applicability to more chaotic physical systems and the fidelity of very long-horizon predictions remain open questions. Nevertheless, these principles offer a promising blueprint for embodied tasks, from robotics to automated agricultural systems. As the field advances toward more capable embodied agents, we advocate embracing diverse generative architectures where data-efficient, structurally-informed models like PGGANs complement large-scale approaches in addressing real-world challenges across domains.

References

Sandra Aigner and Marco Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans, 2018. URL https://arxiv.org/abs/1810.01325.

Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Video generative adversarial networks: A review, 2020. URL https://arxiv.org/abs/2011.02250.

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

Dijun Chen, Kerstin Neumann, Swetlana Friedel, Benjamin Kilian, Ming Chen, Thomas Altmann, and Christian Klukas. Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis . *The Plant Cell*, 26(12):4636–4655, 12 2014. ISSN 1040-4651. doi: 10.1105/tpc.114.129601. URL https://doi.org/10.1105/tpc.114.129601.

Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets, 2019. URL https://arxiv.org/abs/1907.06571.

Lukas Drees, Laura Verena Junker-Frohn, Jana Kierdorf, and Ribana Roscher. Temporal prediction and evaluation of brassica growth in the field using conditional generative adversarial networks. *Computers and Electronics in Agriculture*, 190:106415, November 2021. ISSN 0168-1699. doi: 10.1016/j.compag.2021.106415. URL http://dx.doi.org/10.1016/j.compag.2021.106415.

Hui Feng, Ni Jiang, Chenglong Huang, Wei Fang, Wanneng Yang, Guoxing Chen, Lizhong Xiong, and Qian Liu. A hyperspectral imaging system for an accurate prediction of the above-ground biomass of individual rice plants. *Review of Scientific Instruments*, 84(9), September 2013. ISSN 1089-7623. doi: 10.1063/1.4818918. URL http://dx.doi.org/10.1063/1.4818918.

Ao Fu, Yi Zhou, Tao Zhou, Yi Yang, Bojun Gao, Qun Li, Guobin Wu, and Ling Shao. Exploring the interplay between video generation and world models in autonomous driving: A survey, 2024. URL https://arxiv.org/abs/2411.02914.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017. URL https://arxiv.org/abs/1704.00028.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. URL https://arxiv.org/abs/2204.03458.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models, 2025a. URL https://arxiv.org/abs/2505.14357.

Yiwen Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The gan is dead; long live the gan! a modern gan baseline, 2025b. URL https://arxiv.org/abs/2501.05441.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018a. URL https://arxiv.org/abs/1710.10196.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018b. URL https://arxiv.org/abs/1710.10196.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020a. URL https://arxiv.org/abs/2006.06676.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020b. URL https://arxiv.org/abs/1912.04958.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation, 2025. URL https://arxiv.org/abs/2501.08316.

Chao Liu and Arash Vahdat. Equivdm: Equivariant video diffusion models with temporally consistent noise, 2025. URL https://arxiv.org/abs/2504.09789.

Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, Ajmal Mian, Mubarak Shah, and Chang Xu. Generative physical ai in vision: A survey, 2025. URL https://arxiv.org/abs/2501.10928.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation, 2024. URL https://arxiv.org/abs/2409.18964.

J.M. Montes, F. Technow, B.S. Dhillon, F. Mauch, and A.E. Melchinger. High-throughput non-destructive biomass determination during early plant development in maize under field conditions. *Field Crops Research*, 121(2):268–273, 2011. ISSN 0378-4290. doi: https://doi.org/10.1016/j.fcr.2010.12.017. URL https://www.sciencedirect.com/science/article/pii/S0378429010003400.

Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift gan for large scale video generation, 2020. URL https://arxiv.org/abs/2004.01823.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators, 2024. URL https://arxiv.org/abs/2410.18072.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Abhinav Sagar. Hrvgan: High resolution video generation using spatio-temporal gan, 2025. URL https://arxiv.org/abs/2008.09646.

Hanno Scharr, Hannah Dee, Andrew P. French, and Sotirios A. Tsaftaris. Special issue on computer vision and image analysis in plant phenotyping. *Machine Vision and Applications*, 27(5):607–609, June 2016. ISSN 1432-1769. doi: 10.1007/s00138-016-0787-1. URL http://dx.doi.org/10.1007/s00138-016-0787-1.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2, 2022. URL https://arxiv.org/abs/2112.14683.

Luis Fernando Sánchez-Sastre, Nuno M. S. Alte da Veiga, Norlan Miguel Ruiz-Potosme, Paula Carrión-Prieto, José Luis Marcos-Robles, Luis Manuel Navas-Gracia, and Pablo Martín-Ramos. Assessment of rgb vegetation indices to estimate chlorophyll content in sugar beet leaves in the final cultivation stage. *AgriEngineering*, 2(1):128–149, 2020. ISSN 2624-7402. doi: 10.3390/agriengineering2010009. URL https://www.mdpi.com/2624-7402/2/1/9.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks, 2015. URL https://arxiv.org/abs/1412.0767.

Zhiyu Yin, Kehai Chen, Xuefeng Bai, Ruili Jiang, Juntao Li, Hongdong Li, Jin Liu, Yang Xiang, Jun Yu, and Min Zhang. Asurvey: Spatiotemporal consistency in video generation, 2025. URL https://arxiv.org/abs/2502.17863.

Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling, 2024. URL https://arxiv.org/abs/2412.01821.

Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyan Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis, 2025. URL https://arxiv.org/abs/2503.20822.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. Is sora a world simulator? a comprehensive survey on general world models and beyond, 2024. URL https://arxiv.org/abs/2405.03520.

A Implementation and Training Details

This appendix provides supplementary details regarding the model architecture, loss function, and training hyperparameters.

A.1 Progressive Training Strategy

Our training process closely follows the progressive growing methodology to ensure stability and high-quality results.

Adding Layers for Increased Resolution: Training begins at a very low resolution (4×4 pixels). As the network learns, new layers are smoothly faded in to double the working resolution. This process involves a "transition phase," where the new layers are treated as a residual block with a weight α that increases linearly from 0 to 1, followed by a "stabilization phase" where the network trains at the new resolution. This incremental approach speeds up and stabilizes training, as the network only needs to learn small refinements at each step.

Training Stability Techniques: To prevent unhealthy competition between the generator and discriminator, we employ several stabilization techniques from the PGGAN framework. This includes element-wise weight scaling in all convolutional layers to equalize the learning speed across the network, and pixel-wise feature vector normalization within the generator to prevent the escalation of signal magnitudes.

A.2 Loss Function Formulation

The WGAN-GP loss with epsilon penalty for **optimizing the discriminator** is defined as:

$$L_D(\boldsymbol{x}, \widetilde{\boldsymbol{x}}, \widehat{\boldsymbol{x}}) = \underbrace{\mathop{\mathbb{E}}_{\widetilde{\boldsymbol{x}} \sim \mathbb{P}_g}[D(\widetilde{\boldsymbol{x}})] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r}[D(\boldsymbol{x})]}_{\text{WGAN loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}} \sim \mathbb{P}_{\widehat{\boldsymbol{x}}}} \left[(\|\nabla_{\widehat{\boldsymbol{x}}} D(\widehat{\boldsymbol{x}})\|_2 - 1)^2 \right]}_{\text{gradient-penalty}} + \underbrace{\varepsilon \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r} D(\boldsymbol{x})^2}_{\text{epsilon-penalty}},$$

where $\mathbb{P}_r$ is the data distribution, $\mathbb{P}_g$ is the model distribution implicitly defined by $\widetilde{\boldsymbol{x}} = G(\boldsymbol{z})$, $\widetilde{\boldsymbol{x}} \sim p(\widetilde{\boldsymbol{x}}), \varepsilon$ is the epsilon-penalty coefficient, and λ is the gradient-penalty coefficient. $\mathbb{P}_{\widehat{\boldsymbol{x}}}$ is implicitly defined, sampling uniformly along straight lines between pairs of points sampled from the data distribution $\mathbb{P}_r$ and the generator distribution $\mathbb{P}_g$. The WGAN-GP loss for optimizing the generator is defined as:

$$L_G(\widetilde{\boldsymbol{x}}) = -\mathop{\mathbb{E}}_{\widetilde{\boldsymbol{x}} \sim \mathbb{P}_g}[D(\widetilde{\boldsymbol{x}})]$$

A.3 Network Architecture

The generator and discriminator architectures follow the PGGAN structure, modified with 3D convolutional layers to process spatio-temporal data. A detailed diagram of the generator at multiple resolutions is shown in Figure 4.

Figure 4: Detailed architecture of the PhenoGAN generator, showing the progressive addition of layers to increase resolution from 4×4 (bottom) to 16×16 (middle) and 64×64 (top).

A.4 Hyperparameter Details

The PhenoGAN model was trained using the Adam optimizer with a regularizer. Table 3 summarizes all training hyperparameters used for PhenoGAN.

A.5 Training Progression Visuals

Figure 5 shows the visual evolution of generated samples for the *Arabidopsis thaliana* dataset. The images illustrate the coarse-to-fine learning process central to the PGGAN methodology. The model begins by generating incoherent 4×4 images at early epochs and gradually resolves into structured,

Table 3: PhenoGAN Training Hyperparameters

Hyperparameter	**Value**
Optimizer	Adam
Learning Rate	0.001 (with decay)
β_1	0.0
β_2	0.99
Gradient Penalty (λ)	10
Epsilon Penalty (ε)	0.001

high-fidelity 128×128 images as higher-resolution layers are added and trained. This visualization provides qualitative evidence of stable training and the model's ability to learn increasingly complex features.

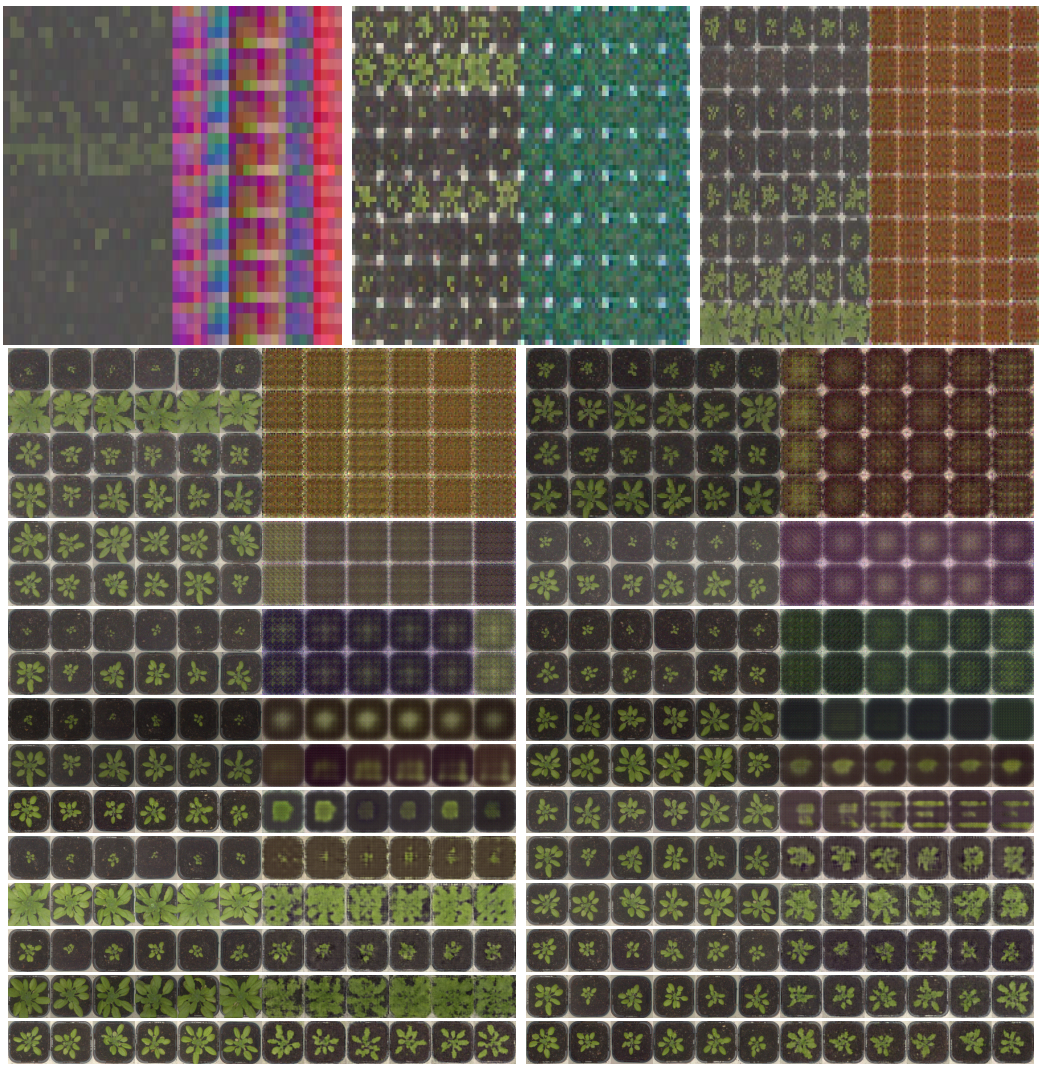

Figure 5: Generated image samples from PhenoGAN throughout training, from epoch 0 to 230. The model's ability to capture structure and detail improves as the resolution doubles from 4×4, to 8×8, and eventually to 128×128.

B Experimental Datasets and Extended Results

B.1 Dataset Details

The experiments were conducted on two publicly available plant phenotyping datasets.

- ***Arabidopsis thaliana*:** This dataset contains approximately 619 images of a model plant species grown under controlled, ideal conditions in a high-throughput environment. Scharr et al. [2016]
- ***Beta vulgaris* (Sugar Beet):** This dataset contains approximately 432 images per condition for sugar beet plants grown under a variety of environmental stressors, including drought, nutrient deficiency, and weed pressure. Sánchez-Sastre et al. [2020]

B.2 Vegetation Index Formulation

The vegetation indices used for evaluating biological plausibility are calculated from normalized RGB chromatic coordinates. First, the R, G, and B channels are normalized:

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B}$$

From these normalized coordinates, the Excessive Green (ExG), Excessive Red (ExR), and Excessive Green-Red (ExGR) indices are calculated as follows:

$$\begin{aligned} \text{ExG} &= 2g - r - b \\ \text{ExR} &= 1.4r - g \\ \text{ExGR} &= \text{ExG} - \text{ExR} \end{aligned}$$

The Projected Leaf Area (PLA) is then calculated by segmenting the image using the ExGR index and counting the relevant pixels.

B.3 Visualization of Vegetation Indices

To provide intuition for the biological plausibility metrics used in the main paper, the following figures show the output of the vegetation index calculations. The ExGR index, in particular, effectively segments the plant biomass from the background soil, forming the basis for the Projected Leaf Area (PLA) calculation. Figure 6 shows the segmentation on both datasets, while Figure 7 shows the ExGR visualization across a temporal sequence.

Table 4: Comparison of Pearson's Coefficient (r) for plant biomass/area prediction against prior work. Our generative approach achieves state-of-the-art performance in capturing physical plausibility for both controlled and stressed plants.

Paper	**Controlled Plant (r)**	**Stressed Plant (r)**
Montes et al. [2011]	0.9517	-
Feng et al. [2013]	0.9675	0.9140
Chen et al. [2014]	0.9891	0.9354
cGAN Drees et al. [2021]	0.877	-
PhenoGAN (ours)	**0.997**	**0.992**

B.4 Per-Condition Performance Analysis

To validate the robustness of PhenoGAN, we analyzed its performance on the *Beta vulgaris* dataset across each of the eight distinct environmental conditions. Figure 8 shows that the model maintains a near-perfect correlation between generated and ground truth vegetation indices across all tested conditions, from "Control" to combined stresses like "Drying - Med N - High Weed."

Furthermore, as shown in Table 4, our model's ability to capture physical plausibility (measured by Pearson's r on segmented plant area) meets or exceeds the performance of prior state-of-the-art methods in plant phenotyping, including those that rely on feature-based approaches rather than direct generative modeling.

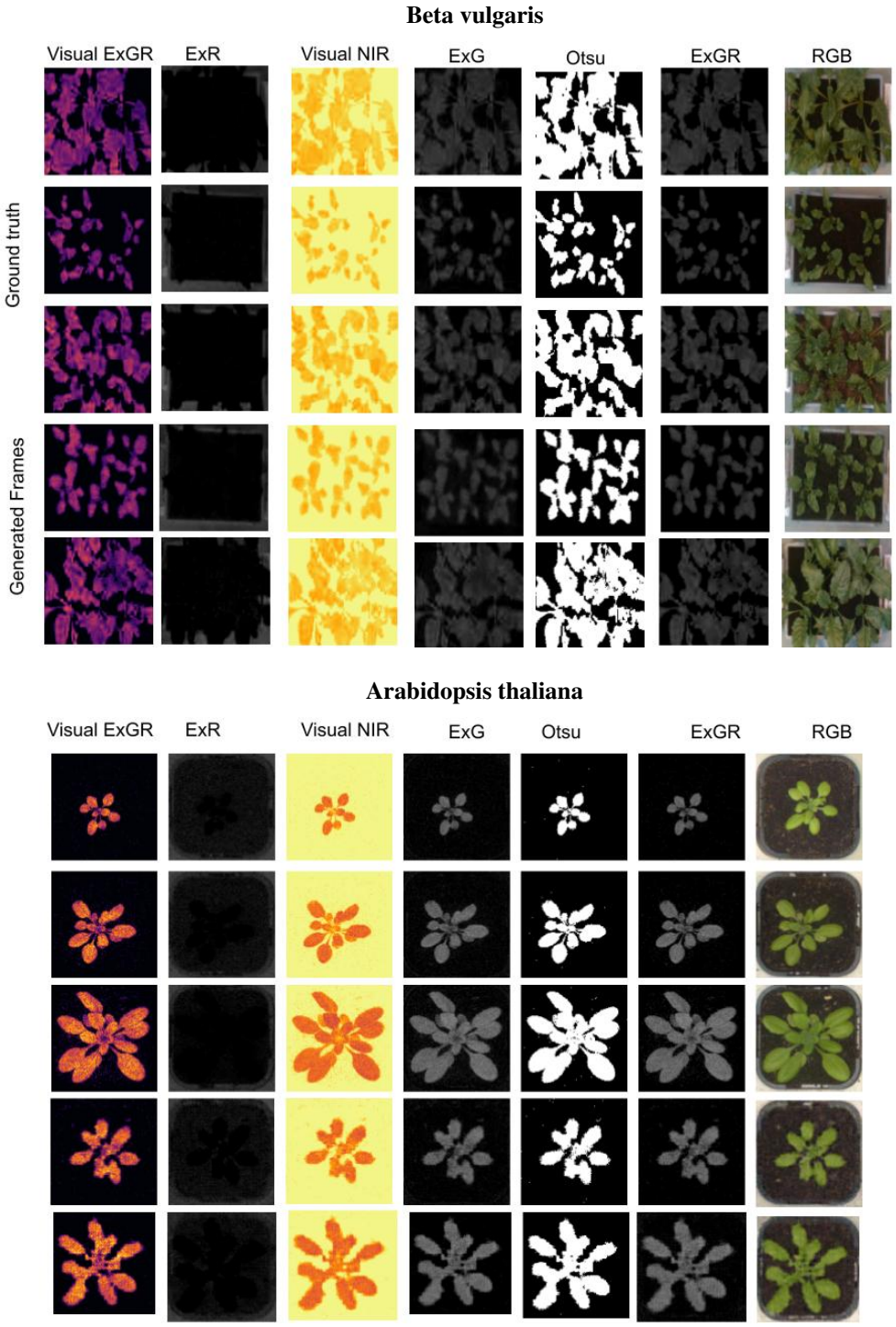

Figure 6: Vegetation indices applied to frames from plant datasets. **Top:** *Beta vulgaris* dataset. **Bottom:** *Arabidopsis thaliana* dataset.

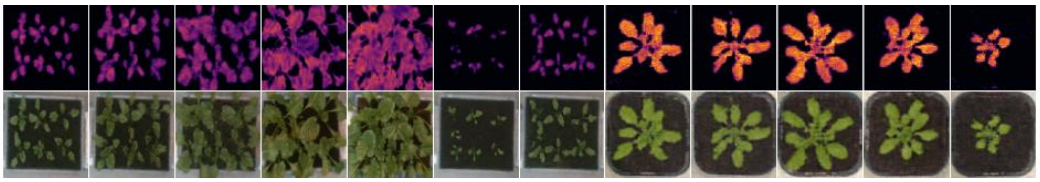

Figure 7: Visual ExGR applied to generated temporal sequences for both datasets.

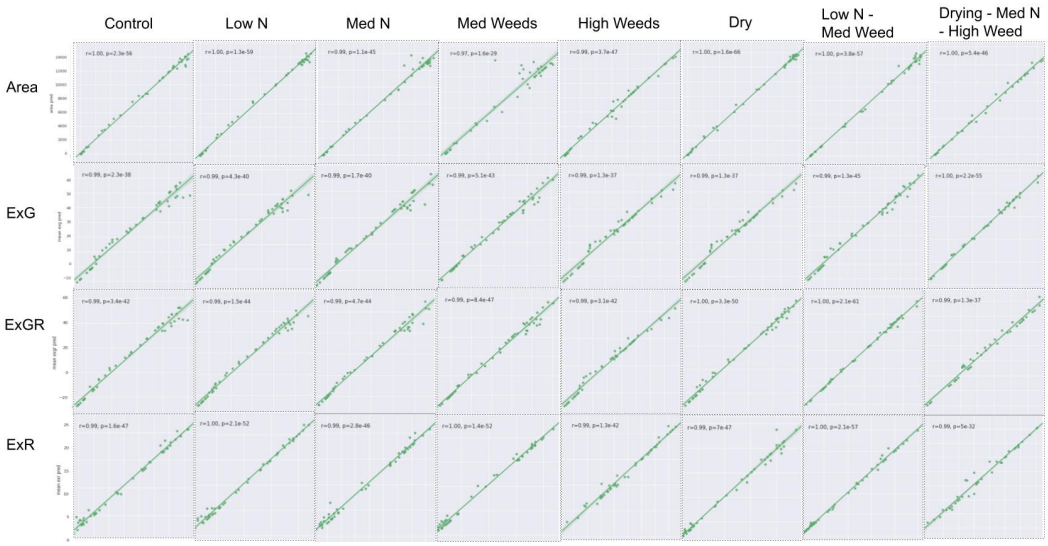

Figure 8: Per-condition correlation plots for the *Beta vulgaris* dataset. The model maintains consistently high correlation across all tested environmental stress conditions, demonstrating its robustness.