EvoVLM: Multimodal Evolutionary Feedback for Visual Symbolic Regression

Hyejin Lee [1] **Junsuk Choe** [1]

Abstract

While Vision-Language Models (VLMs) have demonstrated remarkable capabilities, their potential for visual reasoning in mathematical discovery remains largely underutilized. To address this, we propose EvoVLM, an automated symbolic regression framework that bridges the gap between visual perception and mathematical formalism. By integrating Pruned Exact Linear Time (PELT) segmentation with small VLMs (sVLMs), EvoVLM visually extracts the structural skeleton of data. To refine the equations, we introduce a Multimodal Evolutionary Feedback loop that leverages concurrent textual R^2 metrics and visual overlay plots. Across standard time-series datasets, EvoVLM remains competitive with traditional heuristic methods and, in some cases, outperforms them, notably achieving a higher Best R^2 than Auto-ARIMA and PySR on complex dynamics such as the AirPassengers dataset. By using graph images and summary statistics as the primary inputs to the VLM, while retaining numerical arrays for segmentation and fitness evaluation, EvoVLM establishes a data-efficient and explainable pipeline for visual-driven scientific discovery.

1. Introduction

Recently, there has been a growing demand for explainable modeling techniques to elucidate complex physical phenomena and dynamic systems. While contemporary deep learning models exhibit remarkable predictive performance, their inherent black-box nature severely limits interpretability (Rudin, 2019). Conversely, traditional statistical models offer white-box transparency but often struggle with nonlinearities and distributional shifts, resulting in inconsistent performance across diverse datasets.

To bridge this gap, Symbolic Regression (SR) has emerged as a promising alternative. Conventional methodologies, such as PySR (Cranmer, 2023), provide powerful frameworks for optimizing mathematical structures through numerical search algorithms. Concurrently, recent advanced studies have successfully leveraged Large Language Models (LLMs) and Vision-Language Models (VLMs) for equation discovery (Shojaee et al., 2024; Faroughi et al., 2026), demonstrating the significant utility of utilizing explicit text-based domain descriptions (e.g., physical contexts or variable names) as guiding priors.

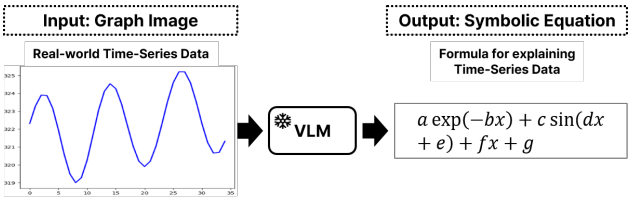


Figure 1. The core concept of Visual Symbolic Regression. Unlike traditional methods relying on raw numerical data, our approach directly leverages VLMs to discover the underlying mathematical equations from visual representations of time-series graphs.

In this paper, we propose a framework that translates real-world time-series data into explainable mathematical formulas in a domain-agnostic manner (see Figure 1). Distinct from setups where prior domain knowledge is readily accessible, our primary motivation is to investigate whether symbolic regression remains feasible when the underlying domain is entirely unknown, relying solely on basic time-series graph images and fundamental summary statistics. To achieve this, our method introduces auxiliary visual indicators directly onto the time-series graphs. By explicitly mapping statistical and geometric properties visually, we enhance the VLM's capacity to directly perceive the underlying structural characteristics from visual representations, thereby providing a complementary and data-efficient alternative for visual-driven scientific discovery.

The main contributions of this paper are summarized as follows:

[1]Department of Data Science and Artificial Intelligence, Sogang University, Seoul, South Korea. Correspondence to: Junsuk Choe <jschoe@sogang.ac.kr>.

Accepted by ICML 2026 AI for Science Workshop.

- **Visual-Guided Mutation Strategy:** We propose a closed-loop visual feedback mechanism where the VLM acts as an evolutionary operator, iteratively correcting geometric discrepancies by visually inspecting formula-data overlay plots.

- **Multimodal Prompting with Statistical Grounding:** Through exhaustive ablation studies across 12 configurations, we systematically align visual context with explicit statistical text prompts to suppress numerical hallucinations.

- **Robustness and Efficiency on Real-World Data:** Utilizing only open-source sVLMs on a single consumer-grade GPU, EvoVLM reliably extracts physically plausible trends from highly non-stationary time-series, bypassing the prohibitive costs of proprietary models and traditional search spaces.

2. Related Work

Symbolic Regression and Heuristic Search. Traditional symbolic regression algorithms aim to discover analytical mathematical expressions that fit given data while maintaining simplicity. State-of-the-art methods like PySR (Cranmer, 2023) have demonstrated strong capabilities in optimizing mathematical structures through genetic programming. However, these purely numerical methods face an exponentially growing search space and require manually predefined basis functions. More critically, when confronted with non-differentiable structural breaks or high-frequency noise in real-world time-series, numerical SR tends to overfit blindly, producing deeply nested and physically implausible equations. EvoVLM bypasses this computational bottleneck by transferring the initial structural reasoning to the visual domain, avoiding brute-force numerical searches.

Vision-Language Models in Scientific Reasoning. The advent of Large Vision-Language Models has extended AI's capability beyond natural language into multimodal scientific reasoning. Recent preliminary works have attempted to use LLMs and VLMs to directly derive mathematical expressions from data arrays or plots (Shojaee et al., 2024; Faroughi et al., 2026). Yet, these approaches primarily treat the model as a single-step function approximator and evaluate performance almost exclusively on synthetic, noiseless physical equations. Consequently, they often suffer from severe structural hallucinations when applied to highly non-stationary distributions. Our work addresses this critical limitation by repositioning the VLM not as a static solver, but as an iterative evolutionary operator, guided by localized segmentation and a closed-loop multimodal feedback system.

3. Methodology

In this section, we detail the EvoVLM framework, designed to discover explainable mathematical equations from real-world time-series data through a three-stage pipeline. Specifically, the system partitions complex data into meaningful segments using the Pruned Exact Linear Time (PELT) algorithm, constructs segment-specific multimodal prompts by combining visual graphs and statistical features, and iteratively optimizes the mathematical expressions via a visual-guided evolutionary search. The overall procedure is summarized in Algorithm 1.

Algorithm 1 EvoVLM: Visual-Semantic Symbolic Discovery

Require: Time-series data Y, Max generations T, Population $N = 4$, Elites $K = 2$
$\{S_1, \ldots, S_m\} \leftarrow$ PELT(Y) {Automated segmentation via changepoints}
$\mathcal{E}_{best} \leftarrow \emptyset$ {Initialize a set to store the best equation for each segment}
for each segment $S \in \{S_1, \ldots, S_m\}$ **do**
  $I_{prompt} \leftarrow$ Construct multimodal visual graph(S)
  $\mathcal{P}_0 \leftarrow$ sVLM(I_{prompt}, Adapters) {Init $N = 4$ base equations via adapter priors}
  for $t = 1, \ldots, T$ **do**
    for each equation $E_i \in \mathcal{P}_{t-1}$ **do**
      $R_i^2 \leftarrow$ Evaluate textual fitness(E_i, S)
      $O_i \leftarrow$ Generate visual overlay plot(E_i, I_{prompt})
    end for
    $\mathcal{P}_{elite} \leftarrow$ Select top $K = 2$ equations based on R_i^2
    $\mathcal{P}_t \leftarrow$ sVLM($\mathcal{P}_{elite}, R^2_{elite}, O_{elite}$) {Visual crossover & mutation}
  end for
  $E_S^* \leftarrow$ Select best equation from final generation $\mathcal{P}_T$
  $\mathcal{E}_{best} \leftarrow \mathcal{E}_{best} \cup \{(S, E_S^*)\}$ {Store the best piecewise equation for current segment}
end for
return Set of piecewise equations $\mathcal{E}_{best}$

3.1. Automated Time-Series Segmentation via PELT

Real-world time-series data often exhibit high volatility and complex, non-stationary dynamics. Directly feeding such lengthy and intricate data into sVLMs can severely degrade their visual reasoning capabilities. To address this, EvoVLM employs the PELT algorithm (Killick et al., 2012) with an RBF (Radial Basis Function) cost model to automatically partition the continuous data into localized, manageable segments based on structural changepoints. This model identifies shifts in both mean and variance by mapping the input signal into a high-dimensional feature space, ensuring robust changepoint detection.

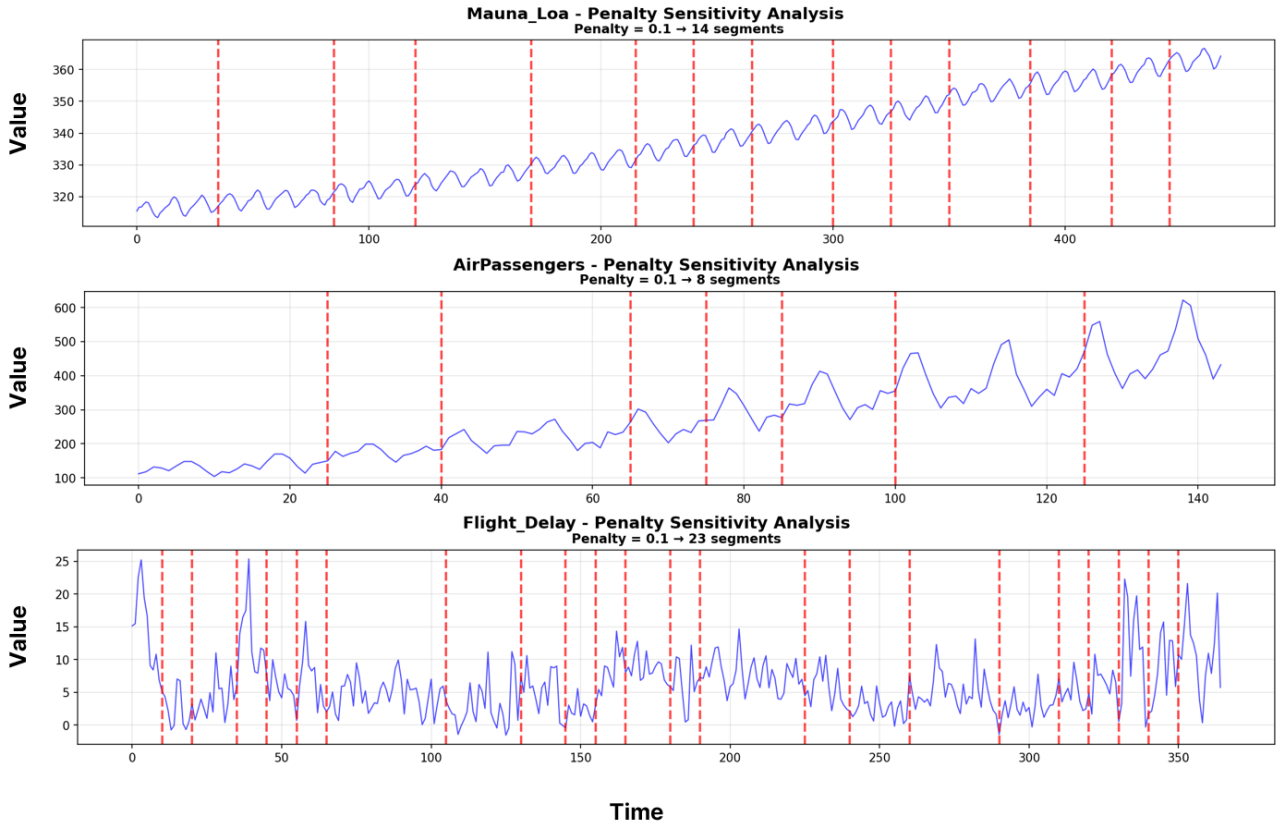


Figure 2. PELT-based segmentation of Mauna Loa, AirPassengers, and Flight Delay datasets with a fixed penalty parameter ($\beta = 0.1$). Vertical red dashed lines indicate changepoints, demonstrating consistent and meaningful partitioning across diverse time-series data characteristics.

Given a time-series $y_{1:T} = (y_1, y_2, \ldots, y_T)$, the PELT algorithm aims to find the optimal number of changepoints m and their positions $\tau_{1:m} = (\tau_1, \ldots, \tau_m)$ by minimizing the following objective function:

$$\min_{m,\tau_{1:m}} \left\{ \sum_{i=1}^{m+1} \mathcal{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta f(m) \right\} \tag{1}$$

where $\mathcal{C}(\cdot)$ represents the empirical cost function (i.e., RBF) for a given segment, β is the penalty parameter, and $f(m)$ is the penalty function to prevent over-segmentation. By integrating this segmentation technique, EvoVLM ensures that each VLM inference step focuses solely on a distinct, homogeneous physical pattern rather than being overwhelmed by global noise.

Crucially, to demonstrate the generalizability and robustness of our framework, we apply a fixed penalty value of $\beta = 0.1$ across all experimental datasets, including the Mauna Loa CO_2 (Keeling et al., 1976), AirPassengers (Box & Jenkins, 1976), and Flight Delay (Ismay & Chunn, 2014) datasets. As illustrated in Figure 2, despite the significant variance in volatility, trend, and seasonality among these datasets, this unified penalty consistently yields semantically meaningful segments. This fixed-parameter approach, bypassing dataset-specific hyperparameter tuning, highlights that EvoVLM is not overly sensitive to parameter changes and establishes a highly robust pipeline for diverse real-world distributions.

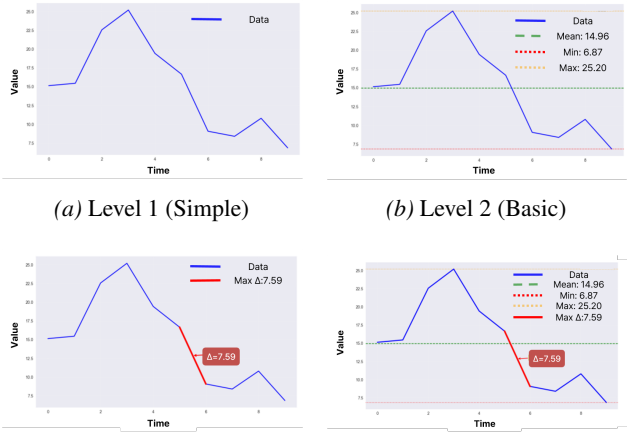
3.2. Multimodal Prompting Strategy

To mitigate hallucinations in sVLMs and maximize their understanding of complex time-series dynamics, we reconstruct the input data across three dimensions: structural context, visual enhancement, and statistical guidance.

Structural Contextualization via Segmentation: Compressing long-term time-series into a single image often leads to significant information loss and visual clutter. By leveraging the PELT-based segments described in Section 3.1, EvoVLM ensures an optimal visual resolution, allowing the model to clearly identify local periodic patterns

and trends without being overwhelmed by global complexity.

Hierarchical Visual Enhancement: To reduce the cognitive load on the sVLM's visual encoder, we systematically enrich the graph images with auxiliary visual hints across four progressive levels (see Figure 3).



(a) Level 1 (Simple) *(b)* Level 2 (Basic)

(c) Level 3 (Advanced) *(d)* Level 4 (Combined)

Figure 3. **Hierarchical Visual Prompting Strategy:** (a) Raw baseline plot; (b) Statistical boundary lines (mean/max/min); (c) Visual annotation of the maximum local fluctuation (Max Δ); (d) Comprehensive integration of all visual aids.

Statistical Guidance via Textual Prompts: To suppress numerical hallucinations—where the model perceives shapes correctly but misinterprets their scales—we inject explicit summary statistics into the text prompt. This serves as a numerical boundary for generating accurate coefficients in the discovered equations. We define two levels of textual information:

- **Basic Stats:** Fundamental metrics including Mean, Median, Min, Max, and Std.
- **Advanced Stats:** Extended features such as quartiles (Q25, Q75), trend slope/intercept, and autocorrelation (lag-1) to assist the model in identifying complex seasonal and trend components.

3.3. The EvoVLM Framework

While the multimodal prompting strategy (Section 3.2) significantly enhances the VLM's initial understanding, zero-shot symbolic generation may still suffer from structural hallucinations. To overcome this, we propose an iterative optimization pipeline that utilizes the VLM as a multimodal evolutionary operator.

Inspired by Language Model Crossover (LMX) (Meyerson et al., 2023) and Optimization by PROmpting (OPRO) (Yang et al., 2024), our framework leverages the semantic reasoning of VLMs to guide the search trajectory. Specifically, EvoVLM incorporates a *dual-feedback loop* consisting of both quantitative R^2 scores and visual overlay plots. As illustrated in Figure 4, the EvoVLM process is structurally aligned into three main phases:

(1) Phase 1: Initialization. The VLM processes the multimodal prompts to generate a diverse initial population of N candidate base equations and their corresponding coefficients (e.g., linear, polynomial, periodic, exponential, complex bases).

(2) Phase 2: Iterative Optimization. This core evolutionary loop refines the population over a defined number of generations (default $T = 4$, adjustable) through a multimodal search mechanism, consisting of three sub-steps:

- *Fitness Evaluation:* Each equation is evaluated against the ground-truth data. Crucially, the VLM receives *multimodal fitness feedback*: the quantitative R^2 score and a *visual overlay plot* comparing the predicted trajectory with the original signal.
- *Elite Selection:* The system ranks candidates based on R^2 scores and selects the top-K elite parents (default $K = 2$, adjustable) to serve as the basis for the next generation.
- *Visual-Guided Crossover & Mutation:* The VLM acts as an evolutionary operator. By visually inspecting the overlay plots to identify discrepancies (e.g., missed local peaks), the VLM performs crossover by combining logical components of parents, and mutation by applying additive terms to fit residuals.

(3) Phase 3: Termination. This iterative optimization continues until the target performance is achieved or the maximum generation limit (T) is reached. As shown in the rightmost panel of Figure 4, the final output is a compositional and explainable mathematical model constructed additively (e.g., Base Trend + Seasonality + Gaussian Pulse) to fully capture the complex time-series dynamics.

4. Experiments

4.1. Experimental Setup

Datasets. As introduced in Section 3.1, we evaluate our framework on three real-world time-series datasets: Mauna Loa CO_2, AirPassengers, and Flight Delay. These benchmarks were specifically selected to encompass a wide spectrum of temporal dynamics, ranging from strict periodicity and multiplicative seasonality to high-frequency volatility, ensuring a comprehensive evaluation of EvoVLM's robustness.

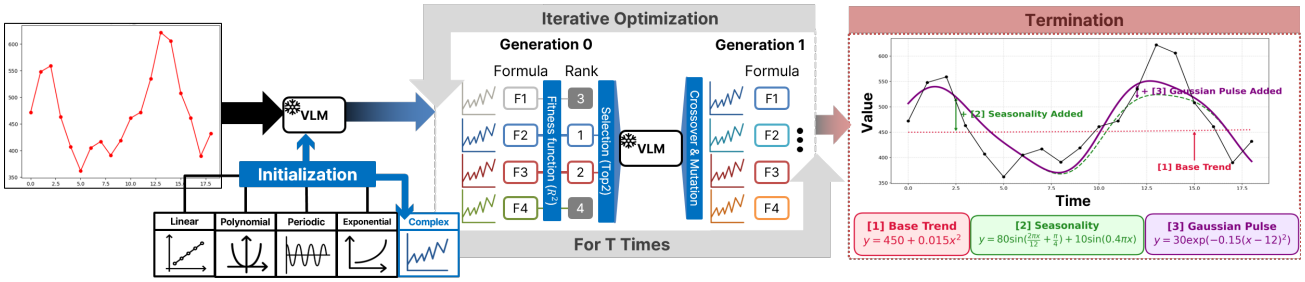


Figure 4. **The Proposed EvoVLM Framework.** The pipeline consists of three main phases: (1) **Initialization**: A VLM generates an initial population of base equations; (2) **Iterative Optimization**: A multimodal evolutionary search refines the equations using visual overlay plots and R^2 fitness scores for selection, crossover, and mutation; and (3) **Termination**: The best formula is constructed additively (e.g., Base Trend + Seasonality + Pulse) to explain the complex time-series dynamics.

Baselines. We compare EvoVLM against a diverse set of representative models: traditional statistical forecasting (**Auto-ARIMA**), standard machine learning approaches (**Linear** and **Polynomial Regression**), and a state-of-the-art search-based symbolic regression framework (**PySR**).

Evaluation Metric. The primary evaluation metric is the Coefficient of Determination (R^2), which quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable. Throughout our experiments, we report the maximum R^2 (hereafter, **Best** R^2) achieved by the generated symbolic expressions.

Implementation Details. Our EvoVLM framework employs the PELT algorithm with a fixed penalty of $\beta = 0.1$ for automated data segmentation. For the evolutionary search loop, the maximum number of generations is set to $T = 4$, with an elite selection size of $K = 2$. Crucially, all experiments, including the multimodal inferences using the Ministral-3 (8B) backbone, were executed on a single consumer-grade NVIDIA GeForce RTX 4060 GPU using PyTorch. This hardware setup highlights the high resource-efficiency and accessibility of our proposed pipeline.

4.2. Main Results: Comparison with Baselines

We evaluate EvoVLM (Ministral-3 backbone, $T = 4$) against statistical and machine learning baselines, with quantitative results summarized in Table 1.

Table 1. Overall performance comparison (Best R^2 ↑) on real-world time-series datasets. The best performing model in each column is highlighted in **bold**, and the second best is underlined.

Model	Mauna Loa	AirPassengers	Flight Delay
Auto-ARIMA	**0.9342**	0.6734	0.5163
Linear Reg.	0.2126	0.2433	0.6297
Poly. Reg.	0.2176	0.4921	0.6507
PySR	0.9222	0.8994	**0.9925**
EvoVLM (Ours)	0.7993	**0.9175**	0.9737

Computational Efficiency. A common criticism of VLM-based agents is their prohibitive inference latency (Table 2).

Table 2. Comparison of computational latency for symbolic generation. For EvoVLM, we report the mean and standard deviation across experimental segments.

Model	Configuration	Latency (s)
PySR	Fixed Search Limit	78.0
EvoVLM (Ours)	Max $T = 4$ Generations	**70.5 ± 7.4**

However, despite executing iterative multimodal inferences up to $T = 4$ generations, EvoVLM achieves an average convergence time of 70.5 ± 7.4 seconds. By not only matching but, on average, slightly outperforming PySR's fixed 78.0-second search limit, EvoVLM definitively proves that advanced visual-semantic search can be highly computationally competitive with standard heuristic-based methods.

Robustness to Non-Stationary Dynamics. EvoVLM demonstrates exceptional adaptability, notably achieving a state-of-the-art Best R^2 of 0.9175 on the AirPassengers dataset. While numerical solvers struggle with its multiplicative seasonality, EvoVLM leverages visual reasoning to explicitly isolate structural shifts into event-driven piecewise functions.

Interpretability vs. Overfitting. While numerical optimizers like PySR excel at achieving a near-perfect R^2 on localized noisy segments (e.g., Flight Delay), this precision often comes at the cost of structural and visual simplicity. As visually demonstrated in Figure 5, fitting non-differentiable structural breaks with a single continuous parametric formula can lead numerical solvers to capture localized noise rather than the underlying physical trend.

To mathematically quantify this trade-off between numerical fit and geometric stability, we evaluate the generated expressions across three dimensions: Abstract Syntax Tree (AST) node counts (k) strictly parsed via the SymPy library (Meurer et al., 2017) to ensure unbiased structural assessment, Bayesian Information Criterion (BIC), and a geometric Roughness penalty ($\int f''(x)^2 dx$) derived from

smoothing splines (Green & Silverman, 1994).

As detailed in Table 3, PySR utilizes 31 AST nodes to maximize the R^2 score, which inadvertently leads to a high Roughness penalty of over 330,000 due to high-frequency oscillations. Conversely, EvoVLM constructs a semantically grounded equation by explicitly decomposing the signal into a piecewise linear macro-trend ($\mathcal{L}(x;K)$) and a localized damped oscillation. By isolating the structural backbone, EvoVLM maintains a competitive statistical fit ($R^2 = 0.9277$, BIC=10.6) while significantly minimizing both AST complexity ($k = 16$) and geometric roughness.

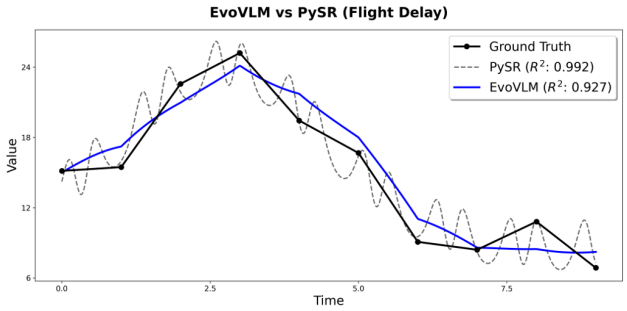


Figure 5. **Qualitative comparison of symbolic regression results.** While PySR (dashed gray line) closely tracks localized noise to maximize numerical precision, EvoVLM (solid blue line) prioritizes the underlying macroscopic trend by decomposing the signal into a structural backbone and a damped oscillation.

Table 3. Quantitative and geometric evaluation on a localized segment of the Flight Delay dataset. While PySR achieves a higher R^2, it results in increased structural complexity and a substantial Roughness penalty. In contrast, EvoVLM significantly reduces the AST complexity (k) and maintains geometric stability.

Model	**k**	**R^2**	**BIC (↓)**	**Roughness (↓)**
PySR	31	0.9925	22.5	330,779
Eq: $y = \frac{\exp(\sin(0.56x)) + \sin(\exp(\dots))}{0.15} + \dots$				
EvoVLM	16	0.9277	10.6	4,098
Eq: $y = \mathcal{L}(x;K) + 2\sin\left(\frac{2\pi x}{5}\right)\exp(-0.1x)$				

Conversely, using visual-semantic representations as the primary interface allows EvoVLM to generate semantically grounded hybrid equations. Rather than forcing a single parametric formula to memorize continuous data points, EvoVLM visually extracts a discrete set of key inflection points (nodes), $K = \{(x_0, y_0), \dots, (x_k, y_k)\}$, to form the geometric skeleton of the sequence.

In our formulation, this skeleton is mathematically realized as a simple piecewise linear interpolant, denoted as $\mathcal{L}(x;K)$. Specifically, for any input x bounded between two adjacent nodes ($x_{i-1} \le x < x_i$), the structural baseline is defined by a standard linear equation:

$$\mathcal{L}(x;K) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}) \tag{2}$$

Unlike idealized physical laws, real-world time-series are frequently governed by discrete external events and non-differentiable structural shifts. By absorbing these macroscopic shocks into a non-parametric linear backbone, the remaining symbolic search elegantly discovers the underlying continuous dynamic forces—such as the localized damped ripple effect ($\sin \times \exp$)—ensuring high physical interpretability without forced curve-fitting.

Furthermore, this explicit isolation of structural nodes (K) transcends mere mathematical curve-fitting; it serves as a powerful diagnostic tool for domain experts. By pinpointing the exact coordinates of macroscopic shocks, EvoVLM provides actionable insights—directing researchers to focus their data exploration on specific timestamps to uncover the real-world events (e.g., policy interventions, system anomalies, or external crises) that triggered the regime shift.

4.3. Backbone Selection: Evaluating sVLMs as Evolutionary Operators

Having demonstrated EvoVLM's superior performance against traditional baselines, we now analyze the critical role of the underlying sVLM. The efficacy of our framework heavily depends on the backbone model's foundational capacity to act as a multimodal evolutionary operator—specifically, its ability to perceive the target data distribution and accurately mutate mathematical expressions.

To ensure a fair and rigorous comparison of their foundational reasoning capabilities, we evaluated all candidate models under identical evolutionary conditions. Each VLM functioned as the mutation operator within the EvoVLM framework, receiving the exact same visual and textual context to refine the mathematical expressions. We evaluated them based on the Best R^2 achieved by the highest-performing formula generated during this standardized process.

Table 4. Performance comparison of various state-of-the-art open-source sVLMs acting as evolutionary operators within our framework. The evaluation is based on the Best R^2 ↑ achieved by the single highest-performing formula for each model under identical experimental conditions.

EvoVLM Backbone	**Params**	**Mauna Loa**	**AirPassengers**	**Flight Delay**
w/ Gemma-3	4B	-0.0545	0.3088	0.5005
w/ Llava-Llama-3	7B	0.0021	0.0567	0.4589
w/ MiniCPM-V	8B	0.1179	0.0444	0.2077
w/ Qwen2.5-VL	7B	0.2821	0.2433	0.6297
w/ Ministral-3 (Ours)	**8B**	**0.7993**	**0.9175**	**0.9737**

As summarized in Table 4, the capacity to perform cross-modal symbolic mutation varies drastically across archi-

tectures. Even under identical standard conditions, models like Qwen2.5-VL and Llava-Llama-3 struggled to map the textual formula to the visual target, often resulting in sub-optimal or collapsed expressions. In stark contrast, **Ministral-3** demonstrated vastly superior and consistent performance across all three diverse datasets.

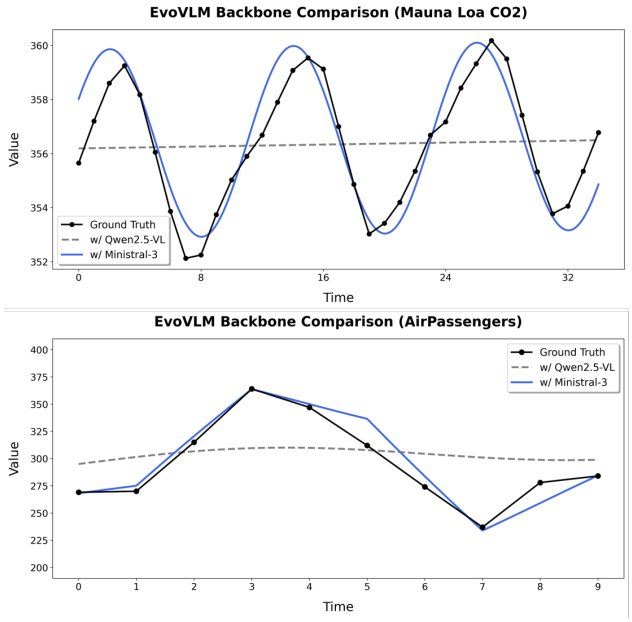


Figure 6. **Qualitative comparison of geometric reasoning during mutation.** The plots contrast the Ground Truth (solid black line) with the generated formulas across two distinct datasets. While Qwen2.5-VL (dashed gray line) collapses to overly smoothed global trends, Ministral-3 (solid blue line) successfully detects critical structural events—accurately recovering complex periodic seasonality (top, Mauna Loa) and sharp non-stationary fluctuations (bottom, AirPassengers).

Beyond quantitative metrics, qualitative analysis of the formulas plotted against the ground truth in Figure 6 reveals a critical distinction in geometric reasoning. When tasked with mutating formulas using standard visual feedback, models like Qwen2.5-VL default to simplistic global trends, such as near-linear or overly smoothed curves, thereby completely missing localized structural events.

In contrast, Ministral-3 demonstrates a sophisticated capacity for visual-to-mathematical translation. By solely observing the identical target image and parent formula, it successfully identifies distinct data behaviors. Specifically, it captures periodic harmonic shifts in the Mauna Loa dataset and critical inflection points like local peaks and deep lows in the AirPassengers dataset. The model then dynamically applies appropriate mathematical operations, including trigonometric additions or piecewise adjustments, to fit these specific segments.

Given its superior multimodal reasoning capabilities, we adopted Ministral-3 as the exclusive backbone architecture for EvoVLM across all subsequent experiments, baseline comparisons, and ablation studies.

4.4. Ablation Study I: Multimodal Prompting Strategies

The core of EvoVLM relies on how visual information is queried and how effectively it is iteratively refined. To rigorously evaluate the individual and synergistic contributions of our design choices, we conducted an exhaustive ablation study comprising 12 distinct configurations. We varied the *Image Prompt* complexity (Simple, Basic, Advanced, Combined) to control the depth of visual context, and the *Stat Prompt* (None, Basic, Advanced) to regulate the level of statistical feedback provided during the iterative search. The Best R^2 results are summarized in Table 5.

Table 5. Ablation study on multimodal prompt combinations. We evaluate four *Image Prompt* complexities against three *Stat Prompt* strategies (None indicates zero-shot generation without statistical feedback). The Best R^2 score for each dataset within this ablation space is **bolded**.

Image Prompt	Stat Prompt	Mauna Loa	AirPassengers	Flight Delay
Simple	None	0.4152	0.2241	0.9435
	Basic	0.6705	0.3014	0.6283
	Advanced	0.1392	0.4819	0.6680
Basic	None	0.1357	0.3089	0.7532
	Basic	0.4720	0.3677	0.8272
	Advanced	0.3847	0.7046	0.9288
Advanced	None	-0.3586	0.5164	0.5471
	Basic	0.3447	0.6149	0.9277
	Advanced	0.4680	0.6897	**0.9737**
Combined	None	0.3477	0.5913	0.7259
	Basic	0.4762	**0.9175**	0.8516
	Advanced	**0.7145**	0.5229	0.8190

Dataset-Dependent Prompt Sensitivity. Rather than revealing a single universal configuration, Table 5 demonstrates that the optimal interplay between visual context and statistical feedback is highly dictated by the intrinsic topology of the target dataset. For instance, the AirPassengers dataset is characterized by macroscopic non-stationary shifts and multiplicative seasonality. Here, providing maximum visual context paired with only moderate statistical grounding (Combined *Image* + Basic *Stat*) yields the optimal Best R^2 of 0.9175. Over-applying statistical feedback in this scenario actually degrades performance, implying that overly strict numerical anchoring can disrupt the VLM's ability to model expanding geometric variance.

Aligning Cognitive Load with Data Characteristics. Conversely, datasets dominated by high-frequency noise and localized shocks, such as Flight Delay, exhibit a completely different prompting dynamic. Providing the model with complex visual instructions without statistical anchoring overwhelms the reasoning engine, dropping performance to 0.5471 (Advanced *Image* + None). To achieve state-of-the-art accuracy (0.9737), the framework requires heavy quantitative constraints (Advanced *Image* + Advanced *Stat*)

to prevent visual hallucinations over the noisy sequence. Interestingly, a purely minimalist approach (Simple *Image* + None) also performs surprisingly well (0.9435) by preventing the model from visually 'over-thinking' the noise entirely. This confirms a crucial finding: effective symbolic discovery requires dynamically aligning the VLM's visual cognitive load and statistical strictness with the specific structural characteristics of the input data.

4.5. Ablation Study II: Evolutionary Trajectory and Convergence

To understand the temporal dynamics of our visual evolutionary search, we analyze the progression of the Best R^2 scores across iterative generations ($T = 0$ to $T = 4$). As illustrated in Table 6, the framework exhibits rapid convergence, though the evolutionary trajectory is highly dataset-dependent.

Table 6. Evolution of the Best R^2 performance across iterative feedback generations (T). $T = 0$ denotes the initial zero-shot prediction.

Dataset	$\mathbf{T=0}$	$\mathbf{T=1}$	$\mathbf{T=2}$	$\mathbf{T=3}$	$\mathbf{T=4}$
Mauna Loa	0.0984	**0.7993**	0.6899	0.7145	0.6972
AirPassengers	-3.5241	0.7804	0.7016	**0.9175**	0.7046
Flight Delay	0.3089	0.9729	0.9728	0.9573	**0.9737**

Overcoming Zero-Shot Collapse. The most striking observation is the catastrophic failure of the unconstrained baseline ($T = 0$) and the subsequent explosive recovery at $T = 1$. In the $T = 0$ state, the VLM operates under extreme informational scarcity: it receives only a minimal visual context (Simple *Image Prompt*) without any **initialization base templates (adapters)** (e.g., linear, periodic, or structural priors). Without these adapter templates to define the valid mathematical search space, the model completely fails to anchor its visual hypothesis to any functional reality, resulting in a severe structural collapse on highly non-stationary datasets like AirPassengers ($R^2 = -3.5241$).

However, entering the first evolutionary critique cycle ($T = 1$) instantly rectifies this macroscopic hallucination, skyrocketing the performance to 0.7804. This definitively highlights a fundamental characteristic of EvoVLM: pure visual reasoning without in-context adapter templates is highly prone to structural hallucination. The massive performance jump at $T = 1$ demonstrates that the iterative feedback loop effectively compensates for this missing structural prior, acting as a powerful numerical anchor that instantly snaps the model's chaotic hypothesis into the correct mathematical space.

Dataset-Specific Convergence Dynamics. The data also reveals that the required depth of evolution heavily depends on the dataset's structural complexity. Datasets with dense, high-frequency noise but stable macro-trends (Flight Delay) or strict periodicity (Mauna Loa) converge almost immediately at $T = 1$ (0.9729 and 0.7993, respectively). Subsequent generations yield only marginal refinements or slight regressions due to over-correction.

In contrast, the AirPassengers dataset—characterized by multiplicative seasonality and expanding structural variance—requires a deeper search. The performance continues to evolve, finally peaking at $T = 3$ (0.9175) as the model iteratively refines the amplitude and decay of its symbolic terms. This dataset-dependent behavior justifies our choice of a maximum generation limit $T = 4$; it provides sufficient evolutionary depth for complex geometric topographies while preventing excessive computational overhead and catastrophic forgetting in simpler distributions.

5. Limitations and Future Work

While EvoVLM establishes a novel visual-symbolic paradigm, its performance is currently constrained by the zero-shot capabilities of the underlying sVLM and faces scaling challenges in visually cluttered multivariate systems. Additionally, our prompting strategies may exhibit backbone-specific brittleness, necessitating broader architectural evaluations. To explicitly isolate the representational benefits of the visual modality, future work will evaluate EvoVLM on comprehensive benchmarks like **LLM-SRBench**, ensuring fair comparisons against textual baselines by incorporating explicit numerical **finite-differences**.

Although Section 4.2 introduces BIC and geometric roughness as post-hoc evaluation metrics, our current evolutionary fitness still relies on numerical R^2. To remove this dependency on raw numerical arrays, future work will explore pure visual-semantic fitness metrics, such as pixel-level Intersection over Union (**IoU**) of curve overlays. By establishing a fully independent visual pipeline, we plan to extend this framework to notoriously difficult non-stationary systems, such as **circadian biology** datasets.

6. Conclusion

In this paper, we proposed **EvoVLM**, a visual symbolic regression framework that bridges the gap between visual perception and mathematical formalism. By explicitly extracting the geometric skeleton of data ($\mathcal{L}(x; K)$) and refining it through a multimodal evolutionary loop, EvoVLM effectively mitigates the numerical overfitting inherent in traditional search-based methods. Our experiments on non-stationary real-world time-series demonstrate that EvoVLM produces structurally compact, semantically grounded, and physically plausible equations, achieving highly competitive accuracy. Ultimately, EvoVLM establishes a highly resource-efficient and cost-effective explainable pipeline,

demonstrating how sVLMs can visually 'understand' data to formulate robust scientific hypotheses.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning by proposing an automated, visually-guided framework for symbolic regression. There are many potential societal consequences of our work, particularly in enabling transparent and explainable AI in scientific discovery, none of which we feel must be specifically highlighted as a negative ethical concern here.

References

Box, G. E. and Jenkins, G. M. *Time series analysis: forecasting and control*. Holden-Day, San Francisco, CA, USA, 1976.

Cranmer, M. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Faroughi, S. A., Mostajeran, F., Arzani, A., and Faroughi, S. Symbolic-KAN: Kolmogorov-arnold networks with discrete symbolic structure for interpretable learning. *arXiv preprint arXiv:2603.23854*, 2026.

Green, P. J. and Silverman, B. W. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1994.

Ismay, C. and Chunn, J. pnwflights14: Data about flights departing PNW airports in 2014. https://github.com/ismayc/pnwflights14, 2014.

Keeling, C. D., Bacastow, R. B., Bainbridge, A. E., Ekdahl Jr, C. A., Guenther, P. R., Waterman, L. S., and Chin, J. F. Atmospheric carbon dioxide variations at Mauna Loa observatory, Hawaii. *Tellus*, 28(6):538–551, 1976.

Killick, R., Fearnhead, P., and Eckley, I. A. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500): 1590–1598, 2012.

Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., Kumar, A., Ivanov, S., Moore, J. K., Singh, S., et al. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017.

Meyerson, E., Nelson, M. J., Bradley, H., Morcos, A. R., Hoover, A. K., and Lehman, J. Language model crossover: Variation through few-shot prompting. *arXiv preprint arXiv:2302.12170*, 2023.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Shojaee, P., Meidani, K., Gupta, S., Farimani, A. B., and Reddy, C. K. LLM-SR: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*, 2024.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024.