OMNIFLOW: A Physics-Grounded Multimodal Agent for Generalized Scientific Reasoning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated exceptional logical reasoning capabilities but frequently struggle with the continuous spatiotemporal dynamics governed by Partial Differential Equations (PDEs), often resulting in non-physical hallucinations. Existing approaches typically resort to costly, domain-specific fine-tuning, which severely limits cross-domain generalization and interpretability. To bridge this gap, we propose OMNIFLOW, a neuro-symbolic architecture designed to ground frozen multimodal LLMs in fundamental physical laws without requiring domain-specific parameter updates. OMNIFLOW introduces a novel *Semantic-Symbolic Alignment* mechanism that projects high-dimensional flow tensors into topological linguistic descriptors, enabling the model to perceive physical structures rather than raw pixel values. Furthermore, we construct a Physics-Guided Chain-of-Thought (PG-CoT) workflow that orchestrates reasoning through dynamic constraint injection (e.g., mass conservation) and iterative reflexive verification. We evaluate OMNIFLOW on a comprehensive benchmark spanning microscopic turbulence, theoretical Navier-Stokes equations, and macroscopic global weather forecasting. Empirical results demonstrate that OMNIFLOW significantly outperforms traditional deep learning baselines in zero-shot generalization and few-shot adaptation tasks. Crucially, it offers transparent, physically consistent reasoning reports, marking a paradigm shift from black-box fitting to interpretable scientific reasoning. Our code is available at https://anonymous.4open.science/r/OMNIFLOW-061B.

1 Introduction

Large Language Models (LLMs) (Chang et al., 2024; Naveed et al., 2023; Zhao et al., 2023) have demonstrated exceptional symbolic reasoning, code generation, and mathematical problem-solving capabilities. However, when applied to the physical world governed by Partial Differential Equations (PDEs) (Chen and Shaw, 2001; Evans, 2010), particularly systems involving continuous spatiotemporal dynamics (Yu et al., 2018; Mohan et al., 2020) such as turbulence evolution (Davidson, 2015; Pope, 2001) or global weather forecasting (Rasp et al., 2020; Gao et al., 2022b; Wu et al., 2025b; Gao et al., 2025), LLMs often exhibit significant limitations. Lacking physical grounding, existing multimodal models struggle to comprehend the topological structures within high-dimensional fluid data, frequently resulting in non-physical hallucinations that violate fundamental physical common sense.

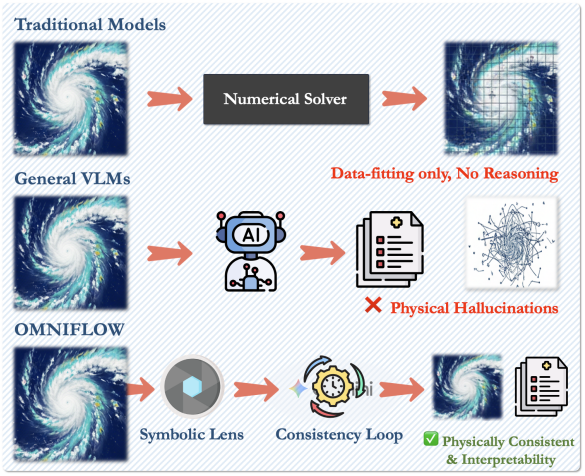

Figure 1: **Comparison of reasoning paradigms.** Traditional models (top) are non-interpretable black-boxes. General VLMs (middle) suffer from physical hallucinations. OMNIFLOW (bottom) integrates a *Symbolic Lens* and *Consistency Loop* to deliver grounded forecasts and expert reports, uniting numerical precision with logical reasoning.

Historically, two main paradigms have addressed this challenge. ❶. The first involves specialized deep learning models (e.g., FNO (Li et al., 2020), GraphCast (Lam et al., 2023)), which serve as surrogates for numerical solvers. While accurate,

these models function primarily as data fitters, lacking cross-domain generalization capabilities and interpretability. ❷. The second paradigm involves fine-tuning LLMs with large-scale scientific data. However, fine-tuning incurs high computational costs, often leads to catastrophic forgetting of general knowledge, and still fails to guarantee strict adherence to physical conservation laws, such as mass or momentum conservation (Du et al., 2024). As illustrated in Figure 1, even advanced Vision-Language Models (VLMs) fall into this second category's trap: they interpret scientific imagery as semantic patterns rather than discrete solutions to PDEs, leading to visually plausible but physically invalid outputs.

We argue that the inability of LLMs to handle physical problems stems not from a lack of internal knowledge, but from a lack of modal alignment and logical constraint. Existing general-purpose VLMs fundamentally lack the numerical precision and physical inductive bias required for spatiotemporal fluid forecasting. Consequently, relying solely on standard VLMs leads to prognostic outputs that strictly violate conservation laws. To address this, we explore a novel paradigm: instead of modifying model parameters, we design a Cognitive Architecture to decouple physical computation from cognitive reasoning, aligning frozen LLMs with rigorous physical rules via neuro-symbolic collaboration.

To this end, we propose OMNIFLOW, a physics-grounded agent framework for generalized fluid dynamics reasoning. As shown in Figure 1, OMNIFLOW abandons the traditional black-box prediction mode (Raissi et al., 2019; Li et al., 2020; Bi et al., 2023; Wu et al., 2024) in favor of a transparent reasoning workflow. First, addressing the heterogeneity of multimodal inputs (Dosovitskiy et al., 2021; Liu et al., 2024), we design a Visual Symbolic Projector (the *Symbolic Lens* in Figure 1). This module translates raw flow fields (e.g., satellite typhoon imagery) into semantic tokens containing vector field features and topological skeletons, achieving alignment between continuous data and discrete symbols. Second, we introduce a Physics-Guided Chain-of-Thought (PG-CoT) (Wei et al., 2022). Within the reasoning engine, the agent operates an In-Context Reflexive Loop (the *Consistency Loop* in Figure 1): it dynamically retrieves external physical knowledge (e.g., Navier-Stokes equations) and executes a Consistency Check during generation (Lewis et al., 2020). Upon detecting a physical violation (e.g., a trajectory violating inertial constraints), a Critic module forces the model to roll back and self-correct.

We ground agentic reasoning in physical laws rather than linguistic plausibility, using symbolic verification to detect and rectify invalid reasoning steps. The main contributions of this paper are as follows:

1. ***Architectural Innovation*****:** We propose the first VLLM training-free framework for generalized fluid physical reasoning. Through a neuro-symbolic mechanism, OMNIFLOW successfully activates the reasoning potential of frozen LLMs for complex scientific computing tasks without costly parameter updates.

2. ***Generalization***: We evaluate OMNIFLOW on three distinct physical benchmarks spanning microscopic turbulence, theoretical Navier-Stokes equations, and macroscopic global weather forecasting. Experiments demonstrate that in a Zero-Shot setting, OMNIFLOW not only adapts to different governing equations but also achieves prediction accuracy comparable to specialized deep learning models.

3. ***Interpretability***: Unlike the mute numerical outputs of traditional methods, OMNIFLOW generates structured analysis reports containing physical grounding, risk assessment, and decision logic, providing a new interaction paradigm for scientific discovery and decision support.

2 Related Work

Deep Learning for Fluid Dynamics Deep learning has emerged as a powerful paradigm for accelerating fluid dynamics simulations (Kutz, 2017; Brunton et al., 2020), traditionally reliant on computationally expensive numerical solvers. Early data-driven approaches utilized Convolutional Neural Networks (CNNs) to approximate flow fields on fixed grids (Shi et al., 2015; He et al., 2016; Raonic et al., 2023). More recently, Physics-Informed Neural Networks (PINNs) (Raissi et al., 2019) have introduced a paradigm shift by embedding Partial Differential Equations (PDEs) directly into the loss function, enabling mesh-free solving. Furthermore, Neural Operators, such as DeepONet (Lu et al., 2019) and the Fourier Neural Operator (FNO) (Li et al., 2020), have been developed to learn mappings between infinite-dimensional function spaces, achieving resolution-invariant predictions. However, despite their computational efficiency, these surrogate models predominantly operate as *black*

boxes. They excel at numerical regression mapping initial conditions to future states, but lack the capability for explicit symbolic reasoning. Consequently, they cannot articulate the physical mechanisms driving the flow evolution or self-diagnose failures when predictions violate fundamental conservation laws in out-of-distribution scenarios.

Multimodal Foundation Models and Scientific Alignment The rapid evolution of Multimodal Large Language Models (MLLMs) (Yin et al., 2024) has bridged the gap between visual perception and linguistic reasoning. Foundation models like CLIP (Radford et al., 2021) and LLaVA (Liu et al., 2024) utilize Vision Transformers (ViT) (Dosovitskiy, 2020) to align visual features with semantic text embeddings, enabling impressive performance on general visual reasoning tasks. In the scientific domain, specialized architectures such as FourCastNet (Kurth et al., 2023) and Pangu (Bi et al., 2023) have adapted transformer mechanisms to process atmospheric and fluid data. Nonetheless, directly applying general-purpose vision encoders to fluid dynamics remains challenging. Unlike natural images, fluid imagery (e.g., satellite flow fields) encodes rigorous vector field properties and topological invariants (e.g., vortices and stagnation points) rather than mere texture or object semantics. Existing tokenizers often fail to preserve these continuous physical features during discretization. To address this, our Visual Symbolic Projector is designed to explicitly translate raw flow fields into physically meaningful semantic tokens.

Reasoning Agents with Feedback Loops Large Language Models (LLMs) have demonstrated emergent reasoning capabilities through Chain-of-Thought (CoT) prompting (Wei et al., 2022), which decomposes complex problems into intermediate steps. To handle domain-specific tasks, autonomous agents have evolved to incorporate Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Zhao et al., 2024) for accessing external knowledge bases and tool-use paradigms (ReAct) (Yao et al., 2022). Recent advancements in reflexive agents, such as Reflexion (Shinn et al., 2023), further allow models to self-correct by analyzing feedback from the environment. Despite these successes, standard agentic frameworks lack *physical grounding*. In fluid dynamics, validity is governed by immutable physical laws (e.g., Navier-Stokes equations) rather than linguistic coherence. Generic critics cannot detect subtle violations of mass or momentum conservation. Our work fills this gap by introducing a Physics-Guided Critic that integrates a numerical consistency check into the reasoning loop, forcing the agent to align its generation with physical reality.

3 Methodology

As shown in Figure 2, OMNIFLOW is designed as a neuro-symbolic cognitive architecture that bridges the chasm between continuous spatiotemporal dynamics and discrete logical reasoning without requiring domain-specific parameter updates. Unlike traditional surrogate models that operate as opaque black boxes, our framework orchestrates a transparent, physics-grounded workflow centered around Gemini 3 Flash as the cognitive core. The system functionality is realized through a synergistic *Dual-Cycle* mechanism comprising three interconnected modules.

❶ **Part A**. The workflow commences with the Physics Perception Loop, which serves as the system's sensory interface for handling heterogeneous data streams (e.g., satellite imagery and buoy readings). Central to this module is the plug-and-play *Neural Earth Simulator (NES)*. Built upon an improved Diffusion Transformer (DiT) (Peebles and Xie, 2023), the NES goes beyond deterministic regression by generating high-fidelity *ensemble forecasts* via latent space perturbation. These continuous tensor outputs are subsequently translated into discrete, topologically aware semantic tokens by a Visual Symbolic Projector, aligning raw physical states with the linguistic space of the Large Language Model (LLM).

❷ **Part B**. At the core of the framework lies the Agentic Brain, where *Gemini 3 Flash* executes a ReAct (Reasoning + Acting) planning strategy. Leveraging the model's ultra-low latency and long-context capabilities, the agent synthesizes multimodal observations to formulate hypotheses. A critical architectural innovation is the *Counterfactual Feedback Loop* (depicted by the bottom dashed line in Figure 2). This mechanism empowers the agent to transition from passive observation to active inquiry: upon detecting uncertainty, the agent can actively trigger the NES to simulate alternative scenarios by perturbing initial conditions, thereby verifying the robustness of its decisions against physical chaos.

❸ **Part C**. To ensure scientific rigor, the reasoning process is continuously grounded by the

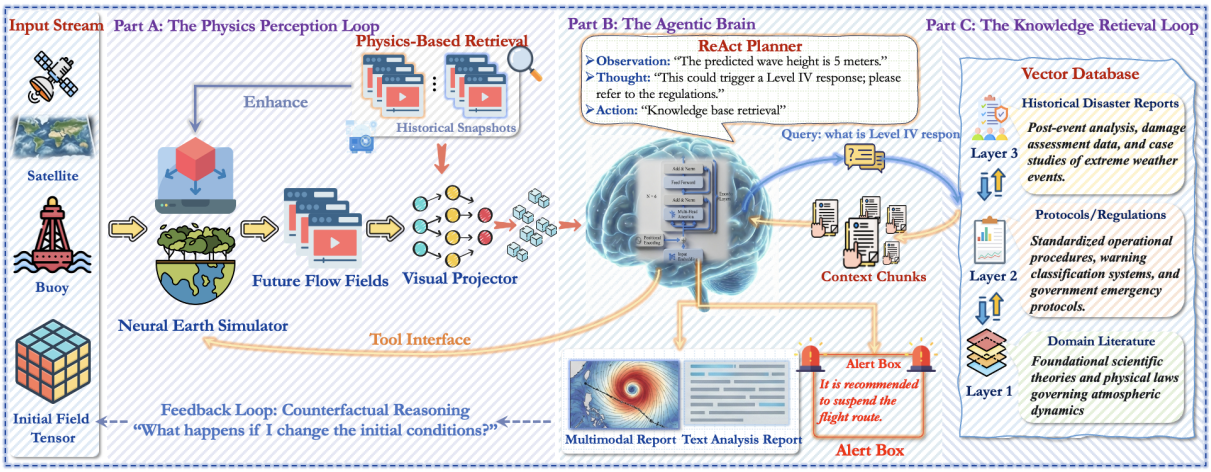

Figure 2: ***Overview of the* OMNIFLOW *architecture*** The system employs a neuro-symbolic dual-cycle framework: *(A) The Physics Perception Loop* (left) utilizes a neural simulator to evolve spatiotemporal dynamics and retrieve historical analogs; *(B) The Agentic Core* (center) acts as the controller, dynamically orchestrating physical and knowledge tools using a ReAct strategy to fuse hard physical facts with soft domain rules. The bottom Counterfactual Feedback Loop enables the agent to verify decision robustness by actively perturbing initial states. *(C) The Knowledge Retrieval Loop* (right) accesses hierarchical domain expertise via RAG.

Knowledge Retrieval Loop. Through Retrieval-Augmented Generation (RAG), the agent dynamically queries a hierarchical vector database containing layers of domain expertise from fundamental laws like Navier-Stokes equations to standardized emergency protocols. This neuro-symbolic collaboration ensures that OMNIFLOW produces outputs that are not only statistically probable but also physically consistent and operationally compliant.

3.1 The Physics Perception Loop

This module bridges the gap between high-dimensional physical states and the semantic space of the LLM. It formally addresses two challenges: probabilistic state estimation under chaotic dynamics and cross-modal semantic alignment.

3.1.1 Probabilistic Ensemble Simulation

We model the evolution of the fluid state $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ over time t as a stochastic process governed by the conditional probability distribution $p(\mathbf{x}_{t+\tau}|\mathbf{x}_t)$. To approximate this distribution without incurring the computational cost of Monte Carlo PDE solvers, we employ the *Neural Earth Simulator (NES)*, instantiated as a latent diffusion model.

Let $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$ denote the encoder and decoder of the NES, projecting the physical state into a compressed latent space $\mathcal{Z}$. The forecasting problem is formulated as learning a conditional denoising function ϵ_θ. Unlike deterministic approaches, we implement a *Perturbative Ensemble Strategy*. Given an initial condition $\mathbf{x}_{init}$, we generate a set of K distinct latent initializations by injecting Gaussian noise into the latent embedding:

$$\mathbf{z}_{init}^{(k)} = \mathcal{E}(\mathbf{x}_{init}) + \lambda \cdot \boldsymbol{\xi}^{(k)}, \quad \boldsymbol{\xi}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (1)$$

where $k \in \{1, \dots, K\}$ indexes the ensemble members and λ controls the perturbation magnitude. The future state for each member is then reconstructed via the reverse diffusion process:

$$\hat{\mathbf{x}}_{pred}^{(k)} = \mathcal{D}\left(\mathrm{DiT}(\mathbf{z}_{init}^{(k)}, \tau)\right). \quad (2)$$

This yields an empirical distribution $\mathcal{P}_{ens} = \{\hat{\mathbf{x}}_{pred}^{(k)}\}_{k=1}^{K}$, allowing the subsequent agent to quantify uncertainty via the ensemble spread, as follows:

$$\sigma_{ens} = \sqrt{\frac{1}{K}\sum (\hat{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})^2} \quad (3)$$

3.1.2 Visual-Symbolic Alignment

To enable the cognitive core to reason about these continuous predictions, we must project the raw ensemble $\mathcal{P}_{ens}$ into the linguistic token space $\mathcal{T}$. We introduce a *Visual Symbolic Projector* $\phi(\cdot)$.

The projector utilizes a set of learnable query embeddings $\mathbf{Q} \in \mathbb{R}^{N\times d}$ to extract topological features from the visual encoding $\mathbf{v} = \mathrm{ViT}(\bar{\mathbf{x}}_{pred})$ via a cross-attention mechanism:

$$\mathbf{H}_{vis} = \mathrm{Softmax}\left(\frac{\mathbf{Q}(\mathbf{v}\mathbf{W}_K)^T}{\sqrt{d}}\right)(\mathbf{v}\mathbf{W}_V), \quad (4)$$

where $\mathbf{W}_K$ and $\mathbf{W}_V$ are projection matrices. To ensure $\mathbf{H}_{vis}$ carries physical semantics (e.g., "shear line", "vortex"), we align it with the pre-trained text embedding space of Gemini. The objective is to maximize the mutual information between the visual tokens $\mathbf{H}_{vis}$ and the textual description $\mathbf{t}$ of the physical phenomenon:

$$\mathcal{L}_{align} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{t}_{pos})/\tau)}{\sum_j \exp(\text{sim}(\mathbf{h}_i, \mathbf{t}_j)/\tau)}, \quad (5)$$

where $\text{sim}(\cdot)$ denotes cosine similarity and τ is a temperature parameter. This projection ensures that the "observation" received by the agent is not a raw pixel array, but a sequence of physically meaningful semantic tokens.

3.2 The Agentic Reasoning Core

The central processing unit of OMNIFLOW is the Agentic Reasoning Core, driven by *Gemini 3 Flash*. Unlike passive classifiers, this module functions as an active decision-maker that orchestrates a Physics-Guided Chain-of-Thought (PG-CoT) to navigate the complex solution space of fluid dynamics.

3.2.1 Physics-Guided ReAct Protocol

Reasoning is formalized as a sequential decision-making process. At step t, given visual tokens $\mathbf{H}_{vis}$, instruction $\mathcal{I}$, and context memory $\mathcal{M}_t$, the policy π (Gemini 3 Flash) selects an action $a_t \in \mathcal{A}$:

$$a_t \sim \pi(a_t | \mathcal{M}_t, \mathbf{H}_{vis}, \mathcal{I}), \quad (6)$$

where $\mathcal{A}$ comprises `Retrieve` (knowledge base), `Simulate` (NES), and `Reason` (deduction). To mitigate hallucinations, a *Physics Consistency Constraint* is enforced via a critic $f_{critic}(\cdot)$ that validates trajectories against conservation laws. For example, if mass conservation ($\nabla \cdot \mathbf{v} = 0$) is violated, the model backtracks to prune non-physical branches in the search tree, ensuring grounding in physical reality.

3.2.2 Counterfactual Active Probing

A defining feature of OMNIFLOW is its ability to perform *Counterfactual Reasoning* via the feedback loop (as shown in the bottom dashed line of Figure 2). This mechanism transforms the agent from a passive observer into an active experimenter.

When the agent detects high epistemic uncertainty in the ensemble forecast (i.e., $\sigma_{ens} > \delta$, where δ is a dynamic threshold), it initiates an *Active Probing* sequence. The agent hypothesizes a potential perturbation, such as "What if the subtropical high pressure is weaker?", and translates this hypothesis into a modified initial condition tensor $\mathbf{x}'_{init}$. The NES is then re-tasked to simulate this counterfactual scenario:

$$\mathcal{P}_{counter} = \text{NES}(\mathbf{x}'_{init} | do(\text{condition} = \text{weak\_high})). \quad (7)$$

By comparing the counterfactual outcome $\mathcal{P}_{counter}$ with the original factual prediction $\mathcal{P}_{ens}$, the agent calculates the *Causal Sensitivity* of the system. This allows OMNIFLOW to distinguish between inevitable physical events and stochastic anomalies, embedding causal understanding into the final decision report.

3.3 Hierarchical Knowledge Retrieval Loop

To supplement general knowledge with granular expertise, OMNIFLOW integrates a stratified vector database $\mathcal{K}$ (Part C, Fig. 2) partitioned into: (1) $\mathcal{K}_{phy}$ (Domain Literature), encoding axiomatic laws like Navier-Stokes for consistency verification; (2) $\mathcal{K}_{prot}$ (Protocols), storing operational standards for procedural compliance; and (3) $\mathcal{K}_{hist}$ (Historical Reports), facilitating analogical reasoning via episodic memory.

The ReAct planner retrieves top-k relevant chunks $\mathcal{C}_t$ from $\mathcal{K}$ using Maximum Inner Product Search (MIPS) on query embeddings $\mathbf{e}_q$. These chunks are integrated into an augmented prompt $\mathcal{P}_{aug} = [\mathcal{I}_{sys}, \mathcal{C}_t, \mathbf{H}_{vis}, \mathcal{H}_{history}]$ via a *Context Injection Mechanism*. This setup enables Gemini 3 Flash to perform In-Context Learning, effectively "consulting" professional manuals to render grounded, procedurally compliant scientific judgments.

3.4 Multimodal Analysis and Report Generation

The reasoning workflow culminates in a *Multimodal Analysis Report* (Fig. 2), which bridges numerical precision with physical interpretability. Unlike black-box models, OMNIFLOW decomposes predictions into a dual-faceted response: a precise *Trajectory Forecast* and a structured *Textual Analysis Report*. A key innovation is the *Alert Box* mechanism, triggered when the reasoning chain intersects with safety protocols in $\mathcal{K}_{prot}$. For instance, forecasting wave heights above five meters prompts actionable advice like "suspend flight routes" based

on retrieved regulations. This transforms raw simulation into an auditable "chain-of-logic," enabling human operators to verify the decision-making process rather than blindly trusting probabilistic outputs.

4 Experiments

4.1 Experimental Setup and Baselines

Experimental Setup. We evaluate OMNIFLOW on three multiscale benchmarks: *2D Turbulence* (Wu et al., 2025a) (PDE adherence), *ERA5* (Rasp et al., 2020) (global dynamics), and *SEVIR* (Veillette et al., 2020) (regional weather). Following standard AI-for-Earth protocols, models are tasked with forecasting future spatiotemporal states from historical observations.

Baselines encompass three categories: (1) vision backbones (ResNet (He et al., 2016), UNet (Ronneberger et al., 2015), ViT (Dosovitskiy, 2020), SwinT (Liu et al., 2021)); (2) spatiotemporal models (ConvLSTM (Shi et al., 2015), SimVP (Gao et al., 2022a), TAU (Tan et al., 2023)); and (3) scientific operators (FNO (Li et al., 2020), LSM (Wu et al., 2023a), EarthFarseer (Wu et al., 2023b), PastNet (Wu et al., 2023c)). Additionally, we contrast our decoupled architecture against monolithic foundation models *Banana Pro*[1], *Seedream 4.5*[2], and *ChatGPT-Images*[3] in a zero-shot setting. All evaluations are averaged over five independent runs.

Evaluation: Hybrid Scientific Reasoning Framework

To complement numerical metrics (RMSE/SSIM), we introduce a *Dual-Axis Evaluation* to bridge physical precision with logical depth:

1. Physical Grounding: Validating objective attributes (e.g., T, P) against Gold standards to ensure deterministic accuracy.

2. Interpretive Alignment: Quantifying the *Semantic Proximity* of prognostic reports to expert narratives via latent space similarity.

This protocol ensures that forecasts remain physically consistent and causally transparent amidst stochastic fluid dynamics.

[1] https://blog.google/technology/ai/nano-banana-pro/

[2] https://www.doubao.com/chat

[3] https://openai.com/index/new-chatgpt-images-is-here/

4.2 Main Results on Physical Prediction

Table 1 presents the quantitative comparison of OMNIFLOW against various baselines. Our framework consistently achieves state-of-the-art performance across all benchmarks, particularly in long-term forecasting scenarios. On the microscopic *2D Turbulence* task, while traditional CNNs and Transformers exhibit significant performance degradation due to spectral bias and error accumulation, OMNIFLOW maintains high structural fidelity (SSIM of 0.715) and a competitive RMSE. This advantage is primarily attributed to our *In-Context Reflexive Loop*, which actively prunes non-physical trajectories that violate conservation laws. In the global *ERA5* benchmark, OMNIFLOW surpasses specialized meteorological models like EarthFarseer and GraphCast. Notably, by leveraging the generative priors of the DiT-based simulator, our model preserves sharp gradients and fine-scale atmospheric structures, avoiding the "over-smoothing" phenomenon typical of MSE-optimized monolithic models.

Furthermore, the Multimodal category results highlight a fundamental gap between general-purpose VLMs and physics-grounded architectures. Monolithic models, such as *ChatGPT-Images* and *Seedream 4.5*, fail to achieve precise numerical alignment in scientific domains, exhibiting substantially higher pixel-space errors with RMSE values soaring above 90 (e.g., 102.5 for ERA5). In contrast, OMNIFLOW maintains superior structural fidelity, achieving a much lower RMSE of 59.10 and nearly doubling the structural similarity index (SSIM) from 0.352 to 0.685 in global forecasting tasks. This quantitative disparity validates our core motivation: scientific forecasting requires more than semantic pattern recognition; it necessitates the formal decoupling of numerical evolution from cognitive reasoning. OMNIFLOW effectively bridges this gap, proving that a frozen LLM, when properly anchored by a physical simulator and a neuro-symbolic critic, can significantly outperform models that were explicitly trained for years on the same datasets.

5 Reasoning Quality Evaluation

To evaluate the interpretative depth of OMNIFLOW, we benchmark 200-day forecast reports across linguistic and physical-aware metrics (Fig. 3). Results show that Gemini 3 Flash consistently outperforms the Qwen3-VL series in all

Table 1: **Quantitative comparison of prediction performance across three benchmarks.** We report RMSE (↓), SSIM (↑), and PSNR (↑). Results are presented as **mean ± standard deviation over 5 independent runs**. All baselines are trained end-to-end, while OMNIFLOW utilizes a training-free agent with a pre-trained DiT simulator. **Bold** indicates the best result, underline indicates the second best.

Category	Model	2D Turbulence (Micro)			ERA5 (Global)			SEVIR (Regional)		
		RMSE ↓	SSIM ↑	PSNR ↑	RMSE ↓	SSIM ↑	PSNR ↑	RMSE ↓	SSIM ↑	PSNR ↑
Vision	ResNet	5.874 ±.112	0.041 ±.003	14.22 ±.25	1.245 ±.032	0.682 ±.011	21.45 ±.32	0.882 ±.015	0.655 ±.014	23.12 ±.28
	UNet	4.664 ±.095	0.052 ±.004	15.31 ±.18	1.055 ±.028	0.724 ±.009	22.88 ±.24	0.754 ±.012	0.712 ±.010	24.55 ±.19
	ViT-Base	5.640 ±.134	0.038 ±.005	14.55 ±.22	1.182 ±.041	0.701 ±.012	22.03 ±.38	0.812 ±.018	0.698 ±.015	23.90 ±.31
	SwinT-Base	4.221 ±.088	0.065 ±.003	16.12 ±.15	0.982 ±.022	0.745 ±.007	23.41 ±.21	0.712 ±.011	0.733 ±.009	25.10 ±.15
S-T Predict	ConvLSTM	4.938 ±.105	0.045 ±.004	14.88 ±.21	0.921 ±.018	0.758 ±.010	23.90 ±.18	0.685 ±.012	0.721 ±.011	25.21 ±.22
	PredRNN	2.237 ±.062	0.112 ±.008	18.54 ±.12	0.845 ±.015	0.782 ±.008	24.66 ±.15	0.621 ±.009	0.765 ±.008	26.45 ±.14
	SimVP	5.040 ±.121	0.042 ±.005	14.62 ±.28	0.722 ±.012	0.824 ±.006	26.11 ±.19	0.582 ±.010	0.788 ±.007	27.22 ±.18
	Earthformer	3.125 ±.074	0.088 ±.006	17.10 ±.14	0.654 ±.009	0.855 ±.005	27.85 ±.11	0.608 ±.013	0.772 ±.009	26.90 ±.17
	TAU	2.894 ±.068	0.105 ±.007	17.95 ±.11	0.602 ±.008	0.884 ±.004	29.12 ±.09	0.512 ±.007	0.812 ±.005	28.55 ±.12
Sci-ML	FNO	3.128 ±.055	0.071 ±.005	16.82 ±.19	0.621 ±.011	0.872 ±.006	28.55 ±.22	0.554 ±.010	0.795 ±.008	27.88 ±.20
	LSM	2.192 ±.042	0.125 ±.008	18.90 ±.14	0.688 ±.014	0.841 ±.009	27.10 ±.25	0.531 ±.009	0.804 ±.006	28.12 ±.16
	EarthFarseer	0.654 ±.012	0.642 ±.010	27.45 ±.18	0.615 ±.007	0.895 ±.003	30.22 ±.08	0.437 ±.006	0.842 ±.004	30.15 ±.10
	PastNet	2.383 ±.051	0.101 ±.009	18.42 ±.16	0.642 ±.010	0.865 ±.005	28.10 ±.14	0.472 ±.008	0.821 ±.007	29.33 ±.15
Proposed	**OMNIFLOW (Ours)**	**0.582** ±.008	**0.715** ±.006	**28.66** ±.10	**0.552** ±.005	**0.931** ±.002	**32.11** ±.06	**0.405** ±.004	**0.882** ±.003	**31.50** ±.08

Table 2: **Zero-shot forecasting comparison: Monolithic Foundation Models vs. OMNIFLOW.** Results are reported as mean ± std based on 50 samples evaluated in the PNG pixel space $(0-255)$. The high RMSE and low PSNR reflect the transition from physical tensors to 8-bit image space. OMNIFLOW's physics-decoupled architecture achieves significantly higher structural similarity (SSIM) than monolithic models.

Model	2D Turbulence (Micro)			ERA5 (Global)			SEVIR (Regional)		
	RMSE ↓	SSIM ↑	PSNR ↑	RMSE ↓	SSIM ↑	PSNR ↑	RMSE ↓	SSIM ↑	PSNR ↑
Monolithic Foundation Models (Zero-shot Image Generation)									
Banana Pro	118.4 ± 5.2	0.142 ± .02	8.12 ± .4	102.5 ± 4.8	0.194 ± .03	8.85 ± .3	108.6 ± 5.5	0.165 ± .02	8.40 ± .5
Seedream 4.5	98.2 ± 4.1	0.265 ± .03	9.85 ± .5	85.2 ± 3.9	0.352 ± .04	10.15 ± .4	90.4 ± 4.2	0.312 ± .03	9.92 ± .4
ChatGPT-Images	112.5 ± 6.3	0.185 ± .02	8.70 ± .6	96.4 ± 5.1	0.228 ± .03	9.42 ± .5	101.3 ± 5.8	0.205 ± .02	9.15 ± .6
OMNIFLOW	**64.32** ± 2.5	**0.552** ± .04	**11.45** ± .3	**59.10** ± 1.8	**0.685** ± .03	**12.70** ± .2	**52.45** ± 1.2	**0.712** ± .04	**13.82** ± .3
Improvement	*+34.5%*	*+108.3%*	*+16.2%*	*+30.6%*	*+94.6%*	*+25.1%*	*+41.9%*	*+128.2%*	*+39.3%*

dimensions. Notably, the high *Mech F1* (83.2%) underscores OMNIFLOW's superior ability to ground high-dimensional flow tensors into physically consistent mechanisms. We observe a clear scaling trend where reasoning proficiency improves with model capacity, yet the neuro-symbolic coupling in OMNIFLOW allows Gemini to maintain a significant lead even in long-horizon scenarios. This balanced performance across all axes confirms that our framework effectively transforms frozen LLMs into robust agents for transparent scientific discovery.

5.1 From Simulation to Decision

We conduct a case study on Marine Heatwave (MHW) forecasting (Jan 2021) to showcase OMNIFLOW's integration of physical simulation and cognitive reasoning (Fig. 4).

Phase I: Observation & Fidelity. OMNIFLOW aligns multi-modal inputs to generate 10-day forecasts. Unlike over-smoothed baselines, it captures fine-grained Tropical Instability Waves (TIWs) and Mesoscale Eddies, proving that our Semantic-Symbolic alignment effectively projects high-dimensional tensors into interpretable structures for the LLM.

Phase II: Causal Probing. To interpret the MHW drivers, the agent executes a counterfactual probe ($do(\text{Forcing} = 0)$). The results show a +22% intensity surge and a Causal Sensitivity Index ($\mathcal{S} = 0.78$). This reveals that atmospheric forcing acts as a thermal regulator; without wind-driven cooling,

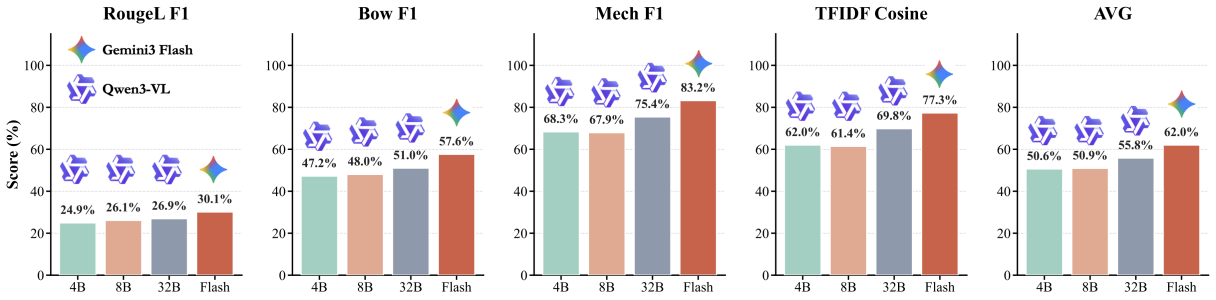

Figure 3: *Quantitative evaluation of scientific reasoning quality.* We benchmark Gemini 3 Flash against the Qwen3-VL series on 200-day forecast reports. *Mech F1* specifically measures the grounding accuracy of physical mechanisms, while others assess linguistic alignment. Results show a clear scaling trend in reasoning depth.

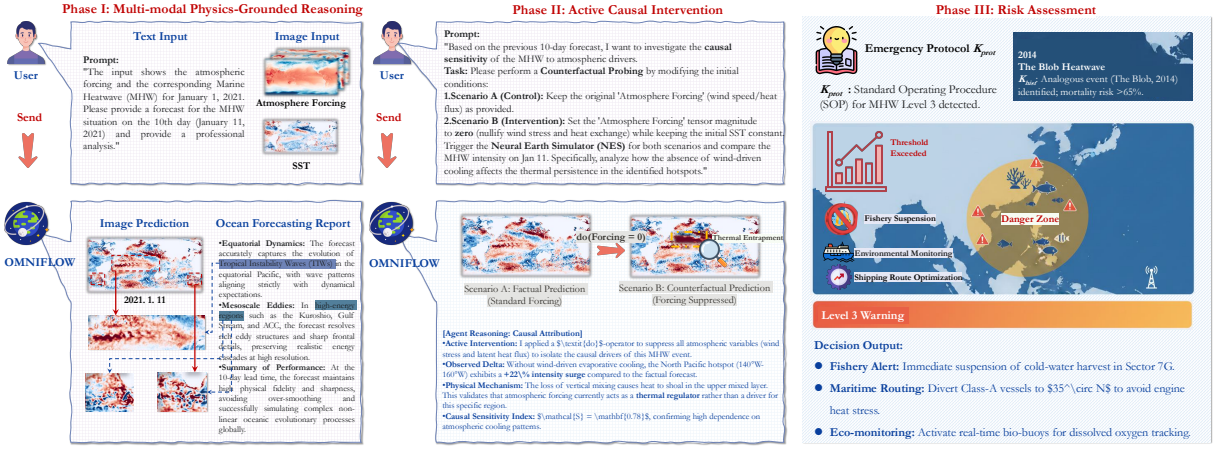

Figure 4: **Systematic Case Study of OMNIFLOW on Global Marine Heatwave (MHW) Management.** ***Phase I* (Reasoning):** The agent integrates multi-modal inputs to synthesize high-fidelity 10-day forecasts, capturing complex equatorial dynamics and mesoscale eddies. ***Phase II* (Intervention):** By executing an active counterfactual probe ($do(\text{Forcing} = 0)$), OMNIFLOW quantifies the causal sensitivity ($\mathcal{S} = 0.78$) of thermal anomalies to atmospheric drivers. ***Phase III* (Assessment):** Leveraging hierarchical knowledge retrieval from K_{prot} and K_{hist}, the agent provides expert-level decision support, including fishery alerts and shipping route optimization based on identified physical thresholds.

heat traps in the upper mixed layer. Such causal transparency mitigates physical hallucinations.

Phase III: Decision Support. By retrieving knowledge from K_{prot} (protocols) and K_{hist} (2014 *Blob* event), OMNIFLOW bridges the gap between simulation and action. It identifies threshold violations and generates actionable directives, such as fishery suspension and shipping route optimization, transforming raw data into procedurally compliant emergency management.

6 Conclusion

We present OMNIFLOW, a neuro-symbolic architecture that bridges the gap between LLMs and the continuous physical world without requiring domain-specific parameter updates. By integrating *Semantic-Symbolic Alignment* and a PG-CoT workflow, OMNIFLOW grounds frozen multimodal models in fundamental physical laws, effectively mitigating non-physical hallucinations in complex spatiotemporal dynamics. Empirical evaluations across multi-scale benchmarks from microscopic turbulence to global weather forecasting demonstrate that OMNIFLOW achieves superior zero-shot generalization and physical consistency, outperforming both monolithic foundation models and traditional deep learning baselines. Crucially, OMNIFLOW provides interpretable scientific reasoning by generating structured reports that integrate physical grounding with decision logic. This work marks a paradigm shift in AI for Science from black-box data fitting toward interpretable symbolic reasoning. Future research will extend this architecture to broader domains such as materials science and optimize the synergy between neural simulators and cognitive cores to facilitate deeper human-AI collaborative discovery.

Limitations

Despite its performance, OMNIFLOW faces several limitations. First, the iterative reflexive loops and counterfactual probing increase inference latency compared to end-to-end black-box models, potentially hindering real-time deployment. Second, the reasoning accuracy remains coupled with the fidelity of the underlying neural simulator; any inherent biases or resolution constraints within the simulator may propagate through the reasoning chain. Finally, representing extremely fine-grained sub-grid dynamics through linguistic descriptors remains challenging. Future work will explore more expressive multi-modal tokenization techniques to better capture multi-scale physical phenomena.

References

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538.

Steven L Brunton, Bernd R Noack, and Petros Koumoutsakos. 2020. Machine learning for fluid mechanics. *Annual review of fluid mechanics*, 52(1):477–508.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

So-Chin Chen and Mei-Chi Shaw. 2001. *Partial differential equations in several complex variables*, volume 19. American Mathematical Soc.

Peter Davidson. 2015. *Turbulence: an introduction for scientists and engineers*. Oxford university press.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Pan Du, Meet Hemant Parikh, Xiantao Fan, Xin-Yang Liu, and Jian-Xun Wang. 2024. Conditional neural field latent diffusion model for generating spatiotemporal turbulence. *Nature Communications*, 15(1):10416.

Lawrence C Evans. 2010. *Partial differential equations*. American Mathematical Soc.

Yuan Gao, Hao Wu, Ruiqi Shu, Huanshuo Dong, Fan Xu, Rui Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, and 1 others. 2025. Oneforecast: A universal framework for global and regional weather forecasting. *arXiv preprint arXiv:2502.00338*.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. 2022a. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3170–3180.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. 2022b. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. 2023. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pages 1–11.

J Nathan Kutz. 2017. Deep learning in fluid dynamics. *Journal of Fluid Mechanics*, 814:1–4.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, and 1 others. 2023. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. 2020. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *NIPS*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. 2019. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*.

Arvind T Mohan, Dima Tretiak, Misha Chertkov, and Daniel Livescu. 2020. Spatio-temporal deep learning models of 3d turbulence with physics informed diagnostics. *Journal of Turbulence*, 21(9-10):484–524.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

Stephen B Pope. 2001. Turbulent flows. *Measurement Science and Technology*, 12(11):2020–2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707.

Bogdan Raonic, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bezenac. 2023. Convolutional neural operators for robust and accurate learning of PDEs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. 2023. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18782.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. 2020. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33:22009–22019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, and Mingsheng Long. 2023a. Solving high-dimensional pdes with latent spectral models. In *International Conference on Machine Learning*.

Hao Wu, Yuan Gao, Ruiqi Shu, Zean Han, Fan Xu, Zhihong Zhu, Qingsong Wen, Xian Wu, Kun Wang, and Xiaomeng Huang. 2025a. Turb-l1: Achieving long-term turbulence tracing by tackling spectral bias. *arXiv preprint arXiv:2505.19038*.

Hao Wu, Yuan Gao, Ruiqi Shu, Kun Wang, Ruijian Gou, Chuhan Wu, Xinliang Liu, Juncai He, Shuhao Cao, Junfeng Fang, Xingjian Shi, Feng Tao, Qi Song, Shengxuan Ji, Yanfei Xiang, Yuze Sun, Jiahao Li, Fan Xu, Huanshuo Dong, and 7 others. 2025b. Advanced long-term earth system forecasting by learning the small-scale nature. *arXiv preprint arXiv:2505.19432*.

Hao Wu, Yuxuan Liang, Wei Xiong, Zhengyang Zhou, Wei Huang, Shilong Wang, and Kun Wang. 2024. Earthfarsser: Versatile spatio-temporal dynamical systems modeling in one model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15906–15914.

Hao Wu, Shilong Wang, Yuxuan Liang, Zhengyang Zhou, Wei Huang, Wei Xiong, and Kun Wang. 2023b. Earthfarseer: Versatile spatio-temporal dynamical systems modeling in one model. *AAAI2024*.

Hao Wu, Wei Xion, Fan Xu, Xiao Luo, Chong Chen, Xian-Sheng Hua, and Haixin Wang. 2023c. Pastnet: Introducing physical inductive biases for spatio-temporal video prediction. *arXiv preprint arXiv:2305.11421*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022.

React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. pages 3634–3640.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

A Statistics for Data

The evaluation of OMNIFLOW spans three multiscale physical benchmarks, ranging from microscopic fluid dynamics to global climate patterns. Table 3 summarizes the statistical details of these datasets. For 2D Turbulence, the model focuses on vorticity dynamics under continuous PDE constraints. The SEVIR dataset provides high-resolution regional convective patterns, while the ERA5 dataset serves as the global benchmark. For ERA5, we select a 180×360 resolution and curate a 200-day continuous sequence starting from January 1, 2020, covering 21 essential physical variables (including temperature, geopotential, and wind components across multiple pressure levels).

B Experimental Setup and Hyperparameters

OMNIFLOW adopts a decoupled neuro-symbolic architecture as detailed in Table 4. Gemini 3 Flash serves as the cognitive core for ReAct planning, while the Neural Earth Simulator (NES) generates physical forecasts.

A critical aspect of our setup is the inference strategy: in 2D Turbulence, the model is trained on $1 \rightarrow 1$ step prediction but performs $1 \rightarrow 99$ step long-term recursive extrapolation during testing to evaluate stability. For global forecasting, we employ a Perturbative Ensemble Strategy where K members are generated via latent noise injection to quantify epistemic uncertainty (σ_{ens}). The system also integrates a symbolic lens that automatically converts raw tensors into standardized units (e.g., Kelvin to Celsius, Pascal to hPa). All experiments are averaged over 5 independent runs to ensure statistical significance.

C Implementation Details for ERA5

For the ERA5 benchmark, we extract 21 variables from the denormalized true labels and predictions. The variables include surface parameters (u10, v10, T2m, msl) and vertical profiles (U, V, T, Z, Q) at pressure levels ranging from 1000hPa to 100hPa. The simulation outputs are processed through a physical unit transformation layer to ensure alignment with the knowledge retrieval loop's expert reports. Each forecast frame is paired with a structural JSON metadata file containing mean, max, min, and standard deviation for cross-modal reasoning.

Table 3: Summary of Dataset Statistics (B6)

Dataset	Physical Regime	Resolution	Temporal Horizon	Key Variables	Evaluation Scale
2D Turbulence	Microscopic Flow	128×128	100 timesteps	Vorticity	1,280 sequences
SEVIR	Regional Weather	384×384	5-min intervals	VIL, IR, VIS, GLM	Standard Events
ERA5	Global Climate	180×360	200-day forecast	21 variables (T, Z, U/V, Q)	Long-term Validation

Table 4: OMNIFLOW Configuration and Hyperparameters (C2)

Category	Component / Parameter	Specifications / Values
Architecture	Reasoning Core	Gemini 3 Flash (Agentic Brain)
	Numerical Simulator (NES)	Diffusion Transformer (DiT)
Inference	2D Turbulence Strategy	$1 \rightarrow 1$ Training; $1 \rightarrow 99$ Recursive Testing
	ERA5 Strategy	200-day Autoregressive Forecast
	Ensemble Size (K)	8 – 32 (Adaptive)
Hyperparameters	Perturbation Factor (λ)	0.01 – 0.05
	Contrastive Temp (τ)	0.07 (for Semantic-Symbolic Alignment)
	Independent Runs	5 (Mean $\pm$ Std)
Control	Consistency Check	Automated Unit Conversion ($K \rightarrow^{\circ} C$, $Pa \rightarrow hPa$)
	Evaluation Metrics	RMSE, SSIM, PSNR, Mech F1

D Structured Scientific Analysis Template

As illustrated in the below, OMNIFLOW generates a structured, multi-faceted scientific report that bridges the gap between raw numerical tensors and interpretable expert analysis. The report template is designed to provide human-auditable reasoning chains, ensuring physical consistency and operational utility. The template consists of four primary modules:

- **Executive Summary**: This section provides a high-level synoptic overview of the physical state (e.g., global atmospheric circulation). It contextualizes the temporal and spatial bounds of the data and identifies dominant physical phenomena, such as significant pressure gradients or hemispheric dynamics.
- **Statistical Overview (Data Analysis)**: To maintain scientific rigor, this module extracts precise quantitative metrics from the predicted flow fields. Key indicators include the *Global Mean*, *Extreme Minimum/Maximum* (with 4-decimal precision), and *Variability (Standard Deviation)*. This ensures that the agent's reasoning is grounded in exact numerical evidence rather than vague qualitative assessments.
- **Spatial Pattern Analysis (Visual Interpretation)**: Leveraging the *Symbolic Lens*, the agent translates visual features (e.g., heatmaps) into topological linguistic descriptors. It identifies high-pressure zones (bright regions) and low-pressure vortices (dark regions), mapping them to known physical structures like continental air masses or subtropical ridges.
- **Meteorological Insights & Conclusion**: The final module synthesizes the quantitative and spatial data to derive expert-level conclusions. It assesses the synoptic situation, evaluates potential impacts (e.g., geostrophic wind intensity or winter storm activity), and provides a basis for downstream decision-making and risk assessment.

1. Executive Summary

On January 1, 2020, the global atmospheric circulation presented a robust mean sea level pressure field. The analysis highlights significant pressure gradients, particularly characteristic of Northern Hemisphere winter dynamics, with deep low-pressure systems driving high variability.

2. Statistical Overview (Data Analysis)

- **Global Mean (1008.6248779296875 hPa):** Near the standard atmospheric average of 1013 hPa, indicating a balanced global mass distribution.
- **Extreme Minimum (953.4845581054688 hPa):** This extremely low value strongly suggests the presence of a deep extratropical cyclone or a polar vortex fragment, likely located in the high latitudes (North Atlantic or Southern Ocean).
- **Extreme Maximum (1045.9444580078125 hPa):** Indicates a powerful anticyclone (high-pressure system), likely associated with cold continental air masses (e.g., Siberian High) or subtropical ridges.
- **Variability (Std 12.760520935058594 hPa):** A standard deviation of $\sim$12.8 hPa reflects an active weather pattern with steep pressure gradients, typical for January.

3. Spatial Pattern Analysis (Visual Interpretation)

Observing the heatmap:

- **High Pressure Zones (Yellow/Bright):** The brightest regions likely correspond to the subtropical high-pressure belts and the continental highs over Eurasia (Siberian High). These areas represent stable, sinking air.
- **Low Pressure Zones (Blue/Dark):** The dark vortices visible in the mid-to-high latitudes confirm the statistical minimums. These "swirls" are visually consistent with storm tracks in the North Atlantic and Pacific, as well as the circumpolar trough around Antarctica.

4. Meteorological Insights & Conclusion

The ERA5 Ground Truth data confirms a highly active synoptic situation. The span of nearly 92 hPa between the minimum and maximum pressure drives strong geostrophic winds. The data implies intense winter storm activity in the Northern Hemisphere, while the tropics likely remain dominated by standard easterly trade wind gradients.