OmniPhys: Knowledge-Graph-Driven Benchmarking and Collective Optimization for Physical Commonsense in Text-to-Image Generation

Yajing Xu
Zhejiang University
Hangzhou, Zhejiang, China
yajingxu@zju.edu.cn

Yarong Lan
Zhejiang University
Hangzhou, Zhejiang, China
yrlan16@zju.edu.cn

Jiaoyan Chen
The University of Manchester
Manchester, United Kingdom
jiaoyan.chen@manchester.ac.uk

Yichi Zhang
Zhejiang University
Hangzhou, Zhejiang, China
zhangyichi.each@zju.edu.cn

Jeff Z. Pan
The University of Edinburgh
Edinburgh, United Kingdom
j.z.pan@ed.ac.uk

Mingchen Tu
Zhejiang University
Ningbo, Zhejiang, China
mingchentz@zju.edu.cn

Zhizhen Liu
Ant Group
Hangzhou, Zhejiang, China
zhizhen.lzz@antgroup.com

Wen Zhang*
Zhejiang University
Hangzhou, Zhejiang, China
zhang.wen@zju.edu.cn

Huajun Chen*
Zhejiang University
Hangzhou, Zhejiang, China
huajunsir@zju.edu.cn

arXiv:2607.25641v1 [cs.CV] 28 Jul 2026

Abstract

While text-to-image models exhibit remarkable visual fidelity, they frequently violate fundamental physical commonsense. Existing benchmarks often rely on coarse-grained descriptions, failing to diagnose the mastery of specific physical principles. Moreover, the high stochasticity of generative processes causes current prompt optimization methods to suffer from "gradient hallucinations," where optimizers are misled by transient visual artifacts rather than systemic flaws. To address these challenges, we introduce **OmniPhys**, a rigorous benchmark of 1,551 samples grounded in a Physical Knowledge Graph. By aligning PhET simulations with standard curricula, OmniPhys operationalizes a knowledge-to-scenario pipeline that performs diagnostic stress tests via a dual-path verification protocol. We further propose **OmniPrompt**, an iterative framework that treats physical alignment as a discrete optimization problem. For each query, OmniPrompt aggregates K stochastic images into a per-query feedback buffer. Across training, it further merges feedback from batches of B queries before each meta-policy update, filtering seed and query-local noise. Evaluations across 12 representative T2I models reveal universal physical bottlenecks. Results demonstrate that OmniPrompt significantly enhances physical consistency across diverse backbones, proving the transferability and efficacy of our evolved meta-policies. The code and data are available at https://github.com/zjukg/OmniPhys

*Corresponding Author

This work is licensed under a Creative Commons Attribution 4.0 International License.
KDD '26, Jeju Island, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817552

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Knowledge representation and reasoning**.

Keywords

Knowledge Graph, Text-to-Image Generation, Physical Commonsense, Collective Optimization

ACM Reference Format:
Yajing Xu, Yarong Lan, Jiaoyan Chen, Yichi Zhang, Jeff Z. Pan, Mingchen Tu, Zhizhen Liu, Wen Zhang, and Huajun Chen. 2026. OmniPhys: Knowledge-Graph-Driven Benchmarking and Collective Optimization for Physical Commonsense in Text-to-Image Generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3817552

1 Introduction

The rapid advancement of Text-to-Image (T2I) models has revolutionized digital content creation, enabling the generation of high-fidelity images from complex natural language descriptions [2, 7, 25–27]. However, a critical "logical ceiling" has emerged as these models transition from artistic tools to simulators of reality: the persistent violation of physical commonsense. As illustrated in Figure 1, even frontier models fail to respect fundamental laws of nature. Examples include unequal liquid levels in communicating vessels, missing refraction-induced displacement at air-water interfaces, and the absence of elastic deformation under mechanical stress. These "physical hallucinations" [23] reveal that current models primarily rely on superficial pattern recognition rather than an internalized understanding of world physics.

Evaluating and rectifying these violations presents two formidable challenges. First, existing benchmarks lack the granularity and scale required for rigorous diagnostics. Early efforts like Commonsense-T2I [10] and PhyBench [19] rely on phenomenological scenarios and provide only coarse-grained analysis at the

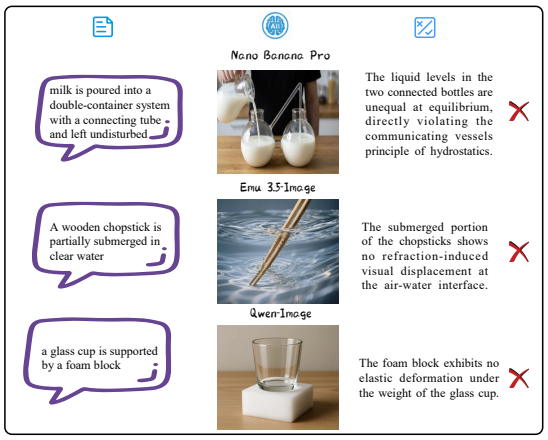

Figure 1: Examples of physical commonsense violations across frontier T2I models, illustrating failures in hydrostatics (top, communicating vessels principle), optics (middle, light refraction), and mechanics (bottom, elastic deformation under stress).

domain level. They may identify general failures in mechanics, but fail to pinpoint specific knowledge gaps. Consequently, they cannot diagnose whether a model lacks mastery of foundational laws, such as Archimedes' principle or the leverage principle. Second, aligning T2I models with physical constraints presents a unique optimization dilemma. Conventional Supervised Fine-Tuning (SFT) [13, 14] and Reinforcement Learning [1, 8, 32] methods are often prohibitively resource-intensive, necessitating both substantial computational overhead and extensive curated datasets of high-fidelity image-text pairs. While discrete prompt optimization provides a lightweight alternative using Large Language Models (LLMs), current methods [24, 33, 34] primarily focus on direct text-to-text interaction. In T2I scenarios, the LLM acts as an intermediary agent. The optimization goal shifts to improving the output of a secondary system (the T2I model) rather than the LLM itself. This setup is plagued by the high stochasticity of the T2I generation manifold. Consequently, the optimizer often suffers from "gradient hallucinations": misleading textual gradients caused by (i) transient visual artifacts from a single stochastic sample and/or (ii) query-local failures that do not generalize across the training pool.

To overcome these hurdles, we first introduce **OmniPhys** (short for **Omni**-category **Phys**ics), a rigorous benchmark comprising 1,551 curated samples anchored to 14 Physical Knowledge Points (PKPs) organized under three domains (Mechanics, Optics, and Object Properties). Its construction operationalizes a **knowledge-to-scenario** pipeline grounded in a hierarchical Physical Knowledge Graph (PKG) [22]. To ensure physical rigor, we align PhET [1] Interactive Simulations with standard physics curricula to extract foundational principles across mechanics, optics, and object properties. Our evaluation employs implicit queries that describe a physical scene without explicitly stating the outcome, such as "an iron block is placed in a container of water". By anchoring each query to a target PKP and its PKG-annotated atomic physical statements, OmniPhys provides a diagnostic "stress test" of a model's authentic reasoning. This design forces the model to derive the correct physical manifestation from its internalized knowledge rather than simple keyword matching. We also implement a Strict **Dual-Path Physical Verification Protocol**, which combines discriminative VQA probes with descriptive consistency auditing. By requiring a model to both correctly identify the physical outcome and maintain structural consistency in its visual execution, this protocol effectively filters out heuristic-based "lucky guesses" and provides a high-fidelity diagnostic of the model's physical alignment.

Building on this diagnostic foundation, we further propose **OmniPrompt**, an iterative framework that transforms physical alignment into a stable, closed-loop optimization process. To mitigate stochastic generation noise, **OmniPrompt** uses a two-level aggregation scheme. At the query level, it consolidates audits over an ensemble of K images into a feedback buffer e_i. At the optimization level, it merges buffers from a batch of B queries before deriving linguistic gradients and updating the meta-policy, so that updates reflect shared physical failures rather than a single query or seed. Consequently, the system evolves high-level instructions that are generalized to diverse physical domains.

In summary, our primary contributions are as follows:

- **Systematic Diagnostics:** We release **OmniPhys**, the first T2I benchmark grounded in a Physical Knowledge Graph. Its taxonomy aligns PhET simulations with standard physics curricula, enabling fine-grained evaluation across 14 PKPs under three domains.
- **Comprehensive Model Analysis:** We conduct an extensive study across 12 representative T2I models, including both frontier closed-source engines and diverse open-source architectures. This provides a granular landscape of current physical reasoning capabilities and identifies universal bottlenecks such as optical reflection and structural mechanics.
- **Robust and Transferable Optimization:** We propose **OmniPrompt**, which addresses "gradient hallucinations" by batch-merging per-query feedback over B training queries before each meta-policy update via textual gradient descent. We demonstrate its efficacy across diverse backbones, proving that our evolved meta-policies provide significant and transferable performance gains.

2 Related Work

2.1 Advancements in Text-to-Image Models

The landscape of T2I synthesis has been recently redefined by a surge in high-fidelity generative capabilities, spanning from stylized art to complex photorealism. Driven by breakthroughs in diffusion processes [2, 7, 29], autoregressive modeling [12, 25] and multimodal unified system [4, 6], modern T2I systems can produce stunning visual content; however, their outputs often betray a lack of physical groundedness. Instead of manifesting a principled understanding of reality, these models frequently default to probabilistic associations, leading to conspicuous failures in maintaining structural integrity or fluidic realism. To bridge this gap, we introduce a novel optimization framework that moves beyond mere visual alignment, iteratively refining generative policies to ensure adherence to essential physical laws.

[1] https://phet.colorado.edu/

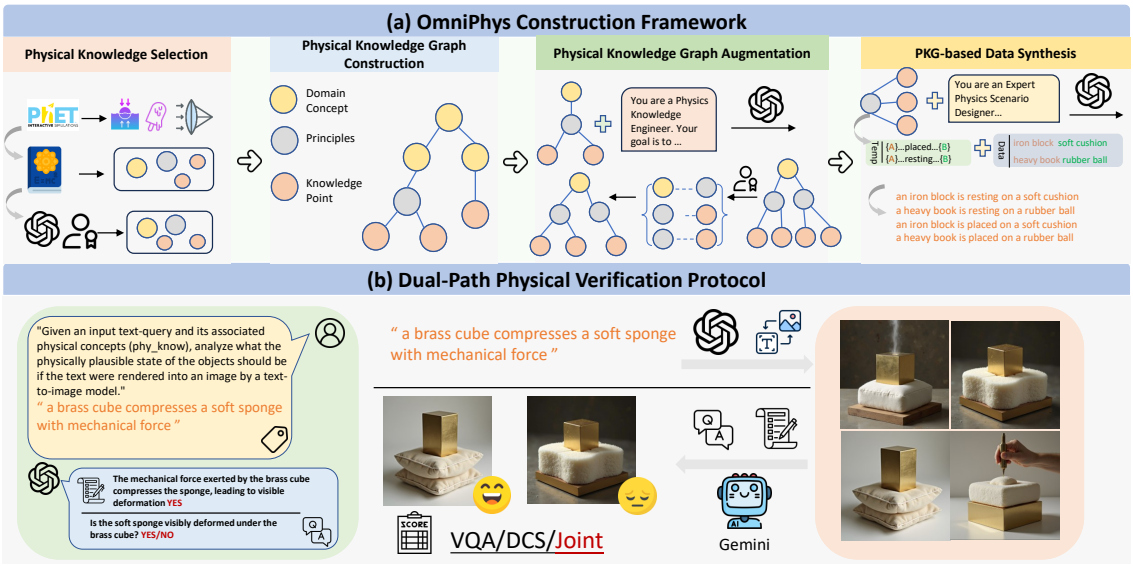

Figure 2: OmniPhys Infrastructure. (a) Construction: A hierarchical pipeline mapping PhET simulations to a PKG for synthesizing high-fidelity prompts. (b) Verification: a dual-path engine that cross-checks discriminative VQA probes against descriptive consistency statements (DCS).

2.2 Evaluation of Text-to-Image Models

Evaluating T2I models remains a multi-faceted challenge. Traditional metrics like FID [16] focus on visual quality, while CLIP-based metrics [15] measure semantic alignment. To capture higher-order capabilities, benchmarks such as ABC-6K [9] and T2I-CompBench [17] have shifted focus toward compositional reasoning, including spatial relations and attribute binding. More recently, WorldGen-Bench [35] and WISE [20] have explored generation grounded in world knowledge, while R2I-Bench [3] targets textual commonsense. Despite these efforts, physical commonsense remains underexplored. Initial attempts like Commonsense-T2I [10] and PhyBench [19] categorize specific physical errors but lack a systematic taxonomy. **Unlike prior work, OmniPhys provides a knowledge-graph-driven taxonomy and a strict dual-path verification protocol for fine-grained physical diagnosis; we further propose OmniPrompt, an iterative meta-policy optimization framework built on this diagnostic signal.**

3 The OmniPhys Benchmark

OmniPhys is a knowledge-graph-driven benchmark designed to address the systematic rigor limitations of existing physical commonsense benchmarks [10, 19]. As illustrated in Figure 2, the benchmark consists of two synergetic modules: (1) a **Construction Framework** that synthesizes high-fidelity, implicit physical queries grounded in a Physical Knowledge Graph; and (2) a **Dual-Path Physical Verification Protocol** that serves as an automated diagnostic engine. Unlike traditional holistic scoring, this protocol decomposes physical consistency into PKG-grounded coupled probes including discriminative visual questions (VQA) and descriptive consistency statements (DCS), thereby ensuring that evaluation is both granular and objective.

3.1 OmniPhys Construction Framework

As illustrated in Figure 2 (a), our pipeline ensures high-fidelity physical alignment and visual verifiability through the following four distinct stages.

(1) Physical Knowledge Selection. To ensure physical rigor, we ground our taxonomy in the PhET Interactive Simulations. We align PhET domains with standard physics curricula to extract principles across mechanics, optics and object properties. Subsequently, we use LLM to filter out abstract concepts (e.g., entropy) that lack unambiguous visual signatures, retaining only those with observable binary states (e.g., sinking vs. floating).

(2) Physical Knowledge Graph Construction. We formalize the curated principles into a PKG, defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which serves as the deterministic backbone for data synthesis and evaluation grounding:

- **Node types $\mathcal{V}$:** *Domain* (e.g., Object Properties), *Domain Concept* (e.g., density), *Principle* (e.g., Archimedes' principle), and leaf *Physical Knowledge Points (PKPs)*. Each PKP anchors one evaluable topic and is annotated with one or more **atomic physical statements**—short, visually verifiable propositions of expected outcomes under the corresponding principle (e.g., float vs. sink vs. suspend under density comparison). These annotations supply PKG-grounded knowledge for synthesis and probe generation; they are not used verbatim as evaluation prompts.
- **Semantic Relations $\mathcal{E}$:** Directed edges typed as `is-a` (Domain → Concept) and `derives` (Concept/Principle → Principle/PKP, or Concept → PKP when no Principle is defined). Atomic statements are stored as node-level annotations on PKPs and are therefore not represented as separate graph nodes or edges.

Full PKG statistics, including the number of PKPs and atomic statements, are reported in Section 3.1.1.

Table 1: PKG coverage by domain. Edge counts include all `is-a` and `derives` relations within each domain.

Domain	Concept	Principle	PKP	Edge	Stmt.
Mechanics	3	2	6	11	9
Object Properties	3	2	3	10	3
Optics	1	5	5	11	5
Total	**7**	**9**	**14**	**32**	**17**

(3) Physical Knowledge Graph Augmentation. To improve the completeness of the PKG in covering visualizable physics, we implement an LLM-based augmentation mechanism. Specifically, using each Domain Concept node as a root, we extract its directly associated Principles and PKPs to form contextual prompts for an LLM. The LLM is tasked with generating candidate PKPs within that domain that are not yet included but remain visually verifiable. Generated candidates are filtered for physical accuracy, logical consistency, and image representability before integration into the graph.

(4) PKG-based Data Synthesis. This stage synthesizes evaluation samples anchored to specific PKPs in the PKG, using the PKP's atomic physical statements as knowledge constraints while deliberately hiding the expected outcome:

- **Scenario Templating:** For each target PKP, LLMs generate diverse scene templates that establish a premise without revealing the result, e.g., *"a {object} is placed into a container of {liquid}"*.
- **Knowledge-Driven Slot Filling:** Based on the material and state constraints encoded in the PKP's atomic statements, compatible entities are selected to create a clear physical contrast (e.g., {object}: iron block; {liquid}: water).
- **Implicit Synthesis:** These components are combined into final implicit prompts that omit the expected physical outcome, such as *"an iron block is placed into a container of water"*. By excluding outcome-revealing keywords (e.g., "sinking"), the prompt forces physically accurate generation to rely on internalized knowledge rather than textual cues.

3.1.1 PKG Scale and Coverage. The finalized PKG is $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = 33$ nodes and $|\mathcal{E}| = 32$ directed edges, organized into four layers—*Domain* (3), *Domain Concept* (7), *Principle* (9), and *PKP* (14)—with 17 atomic physical statements in total (Table 1). Mechanics spans buoyancy (solid–fluid, gas, immiscible liquids), leverage, and pressure (including communicating vessels); Object Properties covers mass–volume–density relations and melting; Optics covers propagation, reflection, refraction, diffraction, and interference. Every PKP satisfies visual verifiability: each atomic statement admits at least one binary observable outcome in image space (e.g., float vs. sink).

PKG-to-benchmark linkage. OmniPhys samples are synthesized from PKG PKPs. Each sample is anchored to one PKP and includes an implicit prompt x, the PKP's atomic physical statements, and LLM-generated coupled probes $(\text{VQA}_i, \text{DCS}_i)_{i=1}^{n}$, where n is sample-specific. The benchmark contains $N = 1{,}551$ implicit prompts in total. All 14 PKPs are represented in the released benchmark (100% PKP coverage).

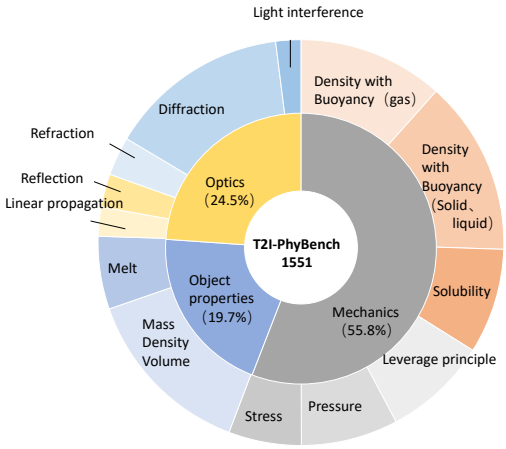

Figure 3: OmniPhys Taxonomy. Hierarchical distribution of 14 PKPs across three domains: Mechanics, Optics, and Object Properties.

3.2 Dual-Path Physical Verification Protocol

Evaluating physical commonsense requires distinguishing between authentic reasoning and accidental artifacts. We propose **Dual-Path Convergence**: an image is deemed physically sound only if it passes (1) a **Discriminative Path** (binary VQA verification) and (2) a **Descriptive Path** (alignment with LLM-generated DCS probes).

(1) PKG-Guided Coupled Probe Generation. For each implicit prompt x, we retrieve the atomic physical statements of its anchor PKP and prompt an LLM with $(x, \{\text{atomic statements}\})$ to generate n coupled units $u_i = (\text{VQA}_i, \text{DCS}_i)_{i=1}^{n}$. Here, n is sample-specific: it denotes the physical state points that the LLM selects to examine for x under PKG constraints, rather than the number of PKG atomic statements. Each unit pairs one discriminative visual question with a matching descriptive consistency statement. Probe generation follows three pillars:

- **Visual-Physical Meta-Template**: Restricts output to observable, static outcomes while pruning transient processes or invisible quantities unjudgeable from 2D projections.
- **Categorized Few-Shot Logic**: Aligns probe generation with the reasoning patterns of the 14 PKG PKPs organized under three domains via expert-curated demonstrations.
- **Coupled Structure**: For each LLM-selected physical state point, the generator co-produces a VQA probe and a matching DCS statement.

Following this PKG-guided pipeline (see Appendix C), the generator outputs structured JSON arrays of coupled $(\text{VQA}_i, \text{DCS}_i)_{i=1}^{n}$ pairs with gold answers.

(2) Automated Scoring and Evaluation. We employ a Vision-Language Model (VLM) as an automated evaluator. For each prompt x_j, we generate 4 independent images $\{G_{j,k}\}_{k=1}^{4}$. Let n_j denote the number of coupled probes generated for prompt x_j. For each image $G_{j,k}$, we adopt a stringent "all-or-nothing" policy:

- **VQA Score**: $S_{VQA}(G_{j,k}) = 1$ if the VLM's responses to all n_j generated discriminative probes match the gold answers; otherwise, 0.

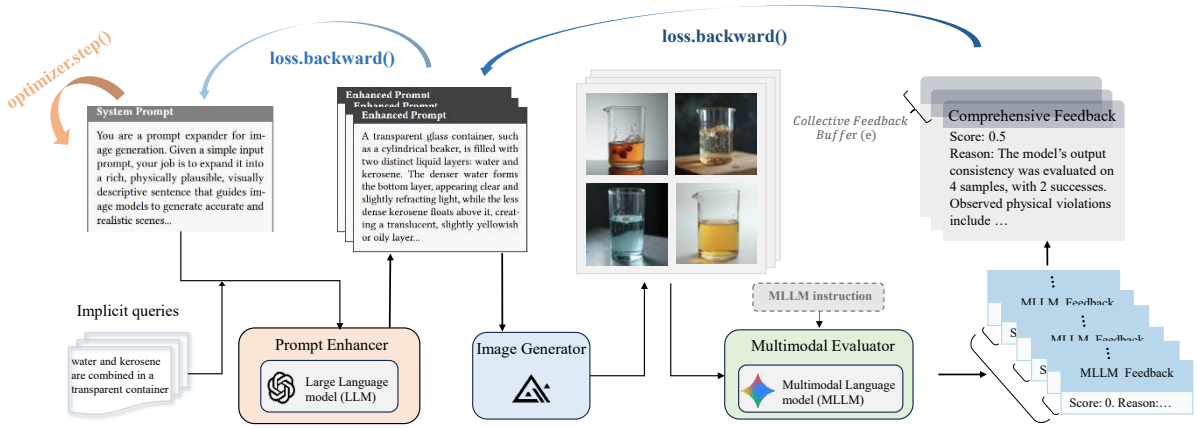

Figure 4: The OmniPrompt Framework. An overview of the iterative alignment loop for physical commonsense in T2I generation. The framework consists of a forward pass for Knowledge-Guided Prompt Enhancement and a backward pass for Meta-Policy Evolution via linguistic gradients.

- **DCS Score**: $S_{DCS}(G_{j,k}) = 1$ if the VLM judges the image to be consistent with all n_j generated descriptive consistency statements; otherwise, 0.
- **Joint Score**: $S_{Joint}(G_{j,k}) = S_{VQA}(G_{j,k}) \times S_{DCS}(G_{j,k})$.

The performance for prompt x_j is the average Joint Score across its 4 generated images:

$$Score(x_j) = \frac{1}{4}\sum_{k=1}^{4} S_{Joint}(G_{j,k})$$

The overall performance on OmniPhys is the mean across all $N = 1,551$ prompts:

$$Score_{Final} = \frac{1}{N}\sum_{j=1}^{N} Score(x_j)$$

This hierarchical scoring ensures that the metric reflects the model's consistent ability to manifest correct physics rather than isolated successes.

4 The OmniPrompt Framework

In this section, we present **OmniPrompt**, an iterative framework that treats physical commonsense alignment as a discrete optimization problem over natural-language instructions. The system establishes a closed-loop pipeline that transforms pixel-level physical violations into structured meta-policy updates through a series of specialized linguistic modules.

4.1 System Formalization and Optimization Goal

We formalize the alignment process as an optimization of the **Meta-Policy** P, which denotes the high-level system instructions that guide prompt transformation. The framework consists of five core components:

(1) **Prompt Enhancer ($\mathcal{L}$):** An LLM-based agent that expands an implicit query x into a physically grounded, visually descriptive prompt $\tilde{x} = \mathcal{L}(x; P)$. The meta-policy P defines the enhancement strategy and is the only component updated during training.
(2) **Stochastic T2I Manifold ($\mathcal{G}$):** A generative module that projects the enhanced prompt $\tilde{x}$ into a high-dimensional visual manifold. We draw an ensemble of samples from $\mathcal{G}(\tilde{x})$ to capture the stochastic distribution of physical manifestations.
(3) **Multimodal Evaluator ($\mathcal{E}$):** A diagnostic module that verifies each generated image against the query's PKG-grounded DCS probes. For every image, $\mathcal{E}$ outputs a binary audit score and a concise rationale, then aggregates them into a collective feedback buffer e_i (Section 4.2.3).
(4) **Feedback Calculator (Γ):** A linguistic derivative operator that propagates batch-level feedback through the TextGrad graph in two stages: it first derives a unified prompt-level gradient $g_{\tilde{x}}$ from the concatenated expansions and feedback buffers, then computes sample-wise meta-gradients $\{g_P^{(i)}\}_{i=1}^{B}$ by backpropagating through each enhancement branch (Section 4.3).
(5) **Meta-Policy Optimizer (Φ):** A batch-aware update operator that rewrites P from the aggregated meta-gradients via Textual Gradient Descent (Section 4.3.2).

Optimization Objective. Let $\mathcal{X}$ denote the training prompt pool used for meta-policy optimization. For each $x_i \in \mathcal{X}$, the enhancer produces $\tilde{x}_i = \mathcal{L}(x_i; P)$, and the evaluator returns a collective feedback buffer e_i with ensemble success rate $\bar{S}_i$ (Section 4.2.3). OmniPrompt seeks the meta-policy

$$P^* = \arg\max_P \ \mathbb{E}_{x_i \sim \mathcal{X}}\left[\bar{S}_i\right]. \tag{1}$$

Only P is updated; the weights of $\mathcal{L}$ and $\mathcal{G}$ remain fixed.

4.2 Physical Synthesis and Diagnostic Evaluation

The forward pass transforms an implicit physical query into a structured diagnostic signal through a sequence of knowledge-driven operations.

4.2.1 Initialization and Reasoning-before-Synthesis ($\mathcal{L}$). To bridge the gap between abstract physics and visual manifestation, we initialize the meta-policy P with a **Reasoning-before-Synthesis** paradigm (initial template in Appendix A). This design ensures that $\mathcal{L}$ infers necessary physical constraints before generating the final description. Specifically, $\mathcal{L}$ identifies governing laws and predicts object interactions to synthesize a detailed, visually grounded prompt $\tilde{x}$.

4.2.2 Multi-sample Visual Synthesis ($\mathcal{G}$). The generative module $\mathcal{G}$ projects $\tilde{x}$ into the visual domain. To mitigate stochastic generation noise, we draw an ensemble of $K = 4$ independent samples $\mathcal{I}_i = \{I_{i,1}, \ldots, I_{i,K}\}$ from $\mathcal{G}(\tilde{x}_i)$ for each enhanced prompt. This multi-sample design provides a stable empirical basis to distinguish systemic reasoning failures from transient artifacts.

4.2.3 Physics-aware Multimodal Evaluation ($\mathcal{E}$). The Multimodal Evaluator ($\mathcal{E}$) serves as a diagnostic bridge between visual pixels and linguistic gradients. Expected physical behaviors are specified by the DCS probes associated with each benchmark query, which are generated from the implicit prompt and PKG atomic statements. Unlike methods that rely on a single holistic score [24], $\mathcal{E}$ first performs per-image audits and then consolidates them into a collective buffer.

Operational Protocol. For each image $I_{i,j} \in \mathcal{I}_i$, $\mathcal{E}$ scrutinizes the alignment between visual manifestations and expected physical behaviors, producing a binary score $s_{i,j} \in \{0, 1\}$ and a concise rationale $r_{i,j}$:

$$s_{i,j} = \begin{cases} 1, & \text{if fully consistent with all DCS probes for } x_i, \\ 0, & \text{if any DCS constraint is violated.} \end{cases} \tag{2}$$

When $s_{i,j} = 0$, the evaluator pinpoints the violation (e.g., "The object's reflection angle is inconsistent with the light source"), providing the raw diagnostic signal for optimization.

Collective Feedback Construction. Per-query aggregation. For each implicit query x_i, $\mathcal{E}$ consolidates individual audits into a collective feedback buffer

$$e_i = \langle \bar{S}_i, \mathcal{R}_i \rangle, \quad \bar{S}_i = \frac{1}{K} \sum_{j=1}^{K} s_{i,j}, \quad \mathcal{R}_i = \bigoplus_{j=1}^{K} r_{i,j}. \tag{3}$$

Here, $\bar{S}_i \in [0, 1]$ is the ensemble success rate over K images, and $\mathcal{R}_i$ is the consolidated rationale. $\bigoplus$ denotes string concatenation.

4.3 Meta-Policy Optimization via Linguistic Gradients

Batch-level aggregation. After processing a batch of B queries through the forward pass, Γ and Φ perform one coupled update on the shared meta-policy P. Only the linguistic modules $\mathcal{L}$ and the meta-policy P participate in the TextGrad computation graph. Image synthesis ($\mathcal{G}$) and multimodal evaluation ($\mathcal{E}$) are external, non-differentiable operators that supply the collective feedback buffer consumed by Γ. The backward pass implements a discrete analog of backpropagation over this linguistic graph.

4.3.1 Feedback Derivation and Joint Differentiation (Γ). The Feedback Calculator (Γ) propagates batch-level feedback through the graph via a two-stage *gradient derivation* procedure.

1. Unified Prompt-level Gradient ($g_{\tilde{x}}$). For batch size B, Γ performs joint batch differentiation over concatenated expansions and feedback buffers:

$$g_{\tilde{x}} \approx \Gamma\left(\bigoplus_{i=1}^{B} \tilde{x}_i \bigoplus \bigoplus_{i=1}^{B} e_i \right). \tag{4}$$

The resulting unified critique is broadcast to each expansion branch via idempotent concatenation backward in the linguistic graph. Specifically, $g_{\tilde{x}}$ pinpoints textual flaws that led to the failures reported by $\mathcal{E}$, including ambiguous support descriptions and missing explanations of buoyancy-related mechanisms.

2. Meta-level Gradient Derivation ($\{g_P^{(i)}\}$). Γ subsequently backpropagates $g_{\tilde{x}}$ to the meta-policy P. For each query i, it analyzes the causal link between the current instruction $P^{(t)}$ and the corresponding expansion failure:

$$g_P^{(i)} = \nabla_P \tilde{x}_i \approx \Gamma\left(P^{(t)}, x_i, \tilde{x}_i, g_{\tilde{x}} \right). \tag{5}$$

Each meta-gradient diagnoses why $P^{(t)}$ failed to guide the enhancer (e.g., missing explicit requirements for structural equilibrium) and provides strategic suggestions for policy refinement.

4.3.2 Meta-Policy Evolution and Update (Φ). The Meta-Policy Optimizer (Φ) refines P via Textual Gradient Descent through a two-stage *policy update* procedure, using the meta-gradients produced by Γ.

Gradient Aggregation. Each meta-gradient $g_P^{(i)}$ is paired with its derivation context and combined into a single textual gradient descent input. This aggregation consolidates batch-level critiques into one update signal and mitigates conflicting sample-wise edits.

Discrete Policy Update. Φ rewrites $P^{(t)}$ in one step to obtain the refined meta-policy:

$$P^{(t+1)} = \Phi\left(P^{(t)}; \{g_P^{(i)}\}_{i=1}^{B} \right). \tag{6}$$

The update is discrete: $P^{(t+1)}$ is a revised natural-language instruction rather than a continuous parameter change. Over successive iterations, P gradually incorporates stronger physical constraints, including mechanistic justification and material consistency, which improves alignment across diverse scenarios.

5 Experiments

5.1 Experimental Setup

5.1.1 OmniPhys Taxonomy. Our benchmark, `OmniPhys`, comprises $N = 1{,}551$ implicit prompts anchored to 14 PKPs from our PKG. As illustrated in Figure 3, samples are organized under three domains:

- **Mechanics (55.8%):** Covers fluid–structure interactions and force equilibrium, including density-driven buoyancy (solid–fluid, gas, and immiscible liquids), leverage, and pressure (including force-induced deformation and communicating vessels).
- **Optics (24.5%):** Evaluates ray and wave optics, including linear propagation, reflection, refraction, diffraction, and interference.

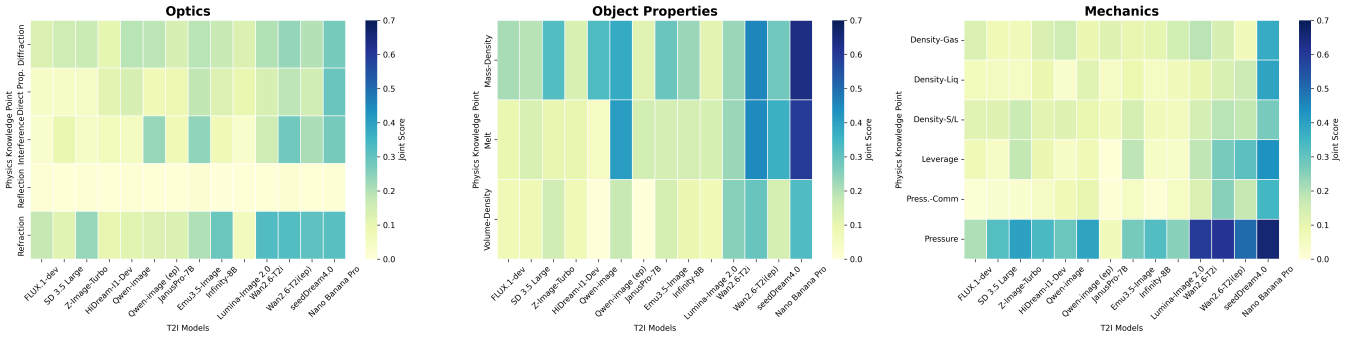

Figure 5: Granular performance across 14 PKPs. Heatmaps show Joint Scores across Optics (left), Object Properties (center), and Mechanics (right). Darker shades represent higher consistency, highlighting systemic bottlenecks like *Reflection* and *Leverage Principle* across all architectures.

Table 2: Performance of 12 T2I models on the `OmniPhys` benchmark (N=1,551 implicit prompts). Scores are reported under the dual-path protocol (VQA, DCS, and Joint). The suffix (ep) denotes the model's native prompt-enhancement configuration (expanded prompt)

Model	Cat.	Params (B)	VQA	DCS	Joint
SD 3.5 Large	Diff	8	0.401	0.158	0.117
FLUX.1-dev	Diff	12	0.391	0.140	0.111
HiDream-l1-Full	Diff	17	0.387	0.153	0.118
Z-Image-Turbo	Diff	6	0.449	0.204	0.156
Qwen-image	Diff	27	0.377	0.167	0.131
Qwen-image (ep)	Diff	27	0.513	0.220	0.176
JanusPro-7B	UM	7	0.348	0.110	0.077
Emu3.5-Image	UM	34	0.469	0.216	0.161
Lumina-Image 2.0	AR	2.6	0.371	0.131	0.102
Infinity	AR	8	0.401	0.156	0.125
SeedDream-4.0	Closed	–	0.514	0.277	0.219
Wan2.6-T2I	Closed	–	0.491	0.266	0.212
Wan2.6-T2I (ep)	Closed	–	0.578	0.348	0.272
Nano Banana Pro	Closed	–	0.572	0.402	0.337

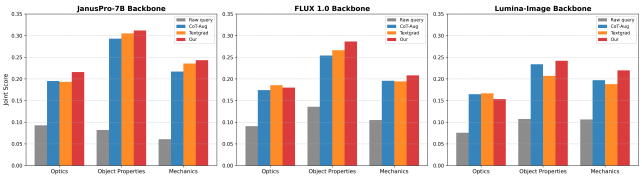

Figure 6: Performance gains across physical domains. Grouped bar charts show Joint Score improvements of `OmniPrompt` over baselines in Optics, Object Properties, and Mechanics.

- **Object Properties (19.7%):** Tests intrinsic material relations, including mass–volume–density consistency and phase change.

5.1.2 Evaluation Landscape: T2I Models. We benchmark our framework against 12 representative T2I models to ensure a robust cross-architecture assessment:

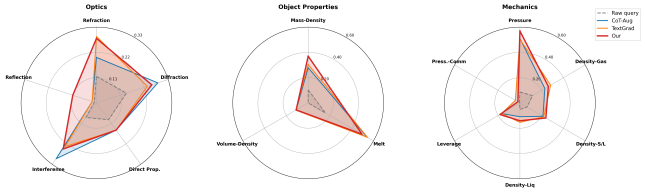

Figure 7: Granular knowledge point recovery on the JanusPro-7B backbone. This radar chart compares the alignment of `OmniPrompt`, TextGrad, and CoT-Aug across 14 PKPs.

- **Open Source**: We select models representing three synthesis paradigms: **Diffusion-based** (FLUX.1-dev [18], SD-3.5-Large [7], Z-Image-Turbo [29], HiDream-I1-Full [2], and Qwen-image [31]), **Unified Multimodal** (JanusPro-7B [4], Emu3.5-Image [6]), and **Autoregressive** (Infinity-8B [12], Lumina-Image-2.0 [25]).
- **Closed source**: We include three leading commercial engines, Wan2.6-T2I [30], SeedDream-4.0 [5], and Gemini 3 Pro Image [11] ("Nano Banana Pro") to represent the industrial upper bound of generative performance.

5.1.3 Baseline Configurations. We establish three baselines to isolate the performance gains from our collective optimization strategy:

- **Raw Query**: Direct generation using original implicit prompts x from OmniPhys.
- **Zero-shot CoT-Aug (CoT-Aug)**: Enhancing prompts x using the initial meta-policy $P^{(0)}$ without any iterative updates.
- **Instance-level Optimization (TextGrad)**: An iterative baseline where the meta-policy is updated using independent, sample-wise gradients.

5.1.4 Implementation Details.

Models and optimization targets. We use GPT-4o (2024-11-20) [21] as the prompt enhancer, TextGrad backward engine, and meta-policy optimizer; Gemini-2.5-Pro [28] as the physics-aware multimodal evaluator. To assess cross-model transfer, evolved policies are evaluated on FLUX.1-dev [18], JanusPro-7B, and Lumina-Image-2.0 [25].

Data splits and training pool. OmniPhys contains 1,551 prompts in total. For experiments, we use a PKP-balanced subset of 1,387 prompts (train/val/test = 274/274/839; ratio 1:1:3), obtained by downsampling *diffraction of light* from 224 to 60. Meta-policy optimization runs on a fixed pool of 37 training prompts (∼2–3 per PKP); validation and test are reserved for rollback and final reporting.

Evaluation and optimization protocol. During training, $\mathcal{E}$ provides DCS-only feedback on the 37 prompt pool; validation and test use the full OmniPhys Dual-Path protocol. Each query uses $K = 4$ generated images. TextGrad updates are triggered every B accumulated queries; we study $B \in \{2, 4, 6, 8, 10, 12\}$. After each update, the candidate policy is validated on the full validation split; we revert to the previous best policy if the Joint Score does not improve, and stop after 5 consecutive failures (patience = 5, up to 5 epochs). An example of End-to-end illustration ($B = 2$, $K = 4$) is found in the Appendix E.

5.2 Benchmarking Results on `OmniPhys`

Table 2 evaluates 12 T2I models on the full `OmniPhys` (N=1,551) using VQA, DCS, and Joint scores. We observe a persistent gap where **VQA scores significantly exceed DCS scores**, suggesting that models can pass binary visual probes via simple heuristics but struggle to satisfy the finer descriptive consistency constraints required by DCS. Consequently, we adopt the **Joint Score** as the primary metric to mitigate this evaluation bias. Overall scores remain relatively low even for strong systems, which is consistent with the strict all-or-nothing Joint criterion on implicit physical queries. By synthesizing the macro trends from Table 2 with the PKP-level heatmaps in Figure 5, we derive three key observations regarding the current T2I landscape:

Frontier Models and Domain-Specific Limits. Closed-source models (e.g., **Nano Banana Pro**, **Wan2.6-T2I (ep)**) achieve the strongest overall performance in physical alignment. However, this superiority is highly domain-specific. As shown in Figure 5, while commercial models excel in *Pressure* (> 0.60), they encounter a persistent bottleneck in *Optics*, where even top-tier industrial engines achieve near-zero scores on PKPs such as *Reflection of light*.

Limited Gains from Model Scale. Increased parameter size does not guarantee a breakthrough in physical reasoning. Large-scale models like **Emu3.5 (34B)** show no transformative advantage over mid-sized counterparts. Scaling primarily improves "texture-based" physics (e.g., material surfaces) but fails on PKPs requiring global mechanical constraints, such as *Leverage Principle*, which exhibit consistently low scores across models. This confirms that raw capacity cannot substitute for explicit structural priors in prompt formulation.

Prompt Expansion as a Material Catalyst. Prompt expansion effectively unlocks latent potential by providing explicit physical context. Significant gains are observed in **Object Properties** PKPs (e.g., *Melt*, *Object properties–mass and density*), as evidenced by the performance leap from **Wan2.6-T2I** to its enhanced version **(ep)**. These detailed descriptions of volume and state changes help bridge the gap between abstract semantic intent and physical visual synthesis. However, expansion offers minimal improvement on logic-intensive *Optics* PKPs, supporting our motivation to optimize meta-policies for reasoning-oriented enhancement.

Table 3: Main results of `OmniPrompt` across T2I backbones. Evaluated on the PKP-balanced split.

Backbone Model	Strategy	VQA	DCS	Joint
FLUX.1-dev	Raw Query	0.389	0.132	0.110
	CoT-Aug	0.478	0.252	0.205
	TextGrad	0.465	0.260	0.209
	OmniPrompt (Ours)	**0.478**	**0.281**	**0.221**
JanusPro-7B	Raw Query	0.333	0.100	0.071
	CoT-Aug	0.499	0.305	0.230
	TextGrad	0.522	0.321	0.244
	OmniPrompt (Ours)	**0.523**	**0.322**	**0.254**
Lumina-Image 2.0	Raw Query	0.365	0.126	0.102
	CoT-Aug	0.465	0.243	0.200
	TextGrad	0.469	0.234	0.189
	OmniPrompt (Ours)	**0.472**	**0.266**	**0.214**

5.3 Effectiveness of `OmniPrompt`

Table 3 and Figure 6 summarize the performance of `OmniPrompt` across three representative T2I architectures. By benchmarking against **Raw Query**, **CoT-Aug**, and **TextGrad**, we evaluate the meta-policy's efficacy in addressing complex physical constraints.

Optimization Dynamics and Stability. `OmniPrompt` consistently establishes the highest performance ceiling among compared methods, achieving Joint Scores of 0.221, 0.254, and 0.214 for FLUX.1-dev, JanusPro-7B, and Lumina-Image-2.0, respectively. Comparison with TextGrad highlights the benefit of batch-merged optimization. While single-query, instance-level backward passes are prone to such gradient hallucinations, OmniPrompt aggregates per-query feedback over B queries before updating P, yielding batch-level diagnostics that are less sensitive to seed- and query-local noise (Appendix D). This robustness is particularly evident on Lumina-Image-2.0, where TextGrad's sample-wise sensitivity causes it to underperform even the simpler CoT-Aug baseline.

Physical State Manifestation. Structured physical descriptions (CoT-Aug) significantly improve DCS scores compared to raw queries, notably surging from 0.100 to 0.305 on JanusPro-7B. This suggests that explicit descriptive consistency constraints provide a necessary foundation for models to move from simple object recognition toward physically checkable execution. `OmniPrompt` further refines these descriptions through meta-policy evolution, yielding more consistent manifestations of force and material interactions.

Domain-Specific Knowledge Recovery. As shown in Figures 6 and 7, the evolved meta-policy is particularly effective on **Mechanics** and **Object Properties** PKPs. `OmniPrompt` successfully improves challenging cases such as *Leverage Principle* and *Pressure-communicator*, which simple text expansion fails to resolve. For material-centric PKPs, it guides the model toward more verifiable visual evidence, such as puddle formation for *Melt* or clearer volume–density relations for mass–density PKPs. While *Reflection of light* remains difficult, we observe meaningful progress on other *Optics*

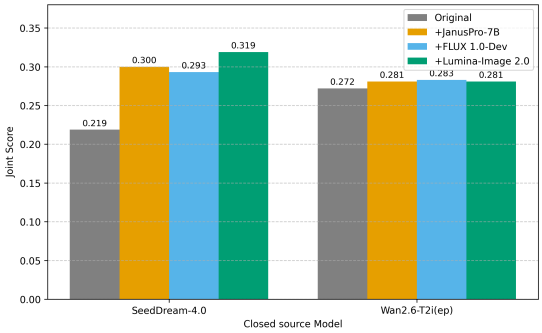

Figure 8: Cross-Model Policy Generalization. Results showing the transfer of meta-policies evolved on open-source backbones to closed-source engines.

Table 4: Human evaluation of VQA-DCS probe quality and VLM scoring alignment. For 50 randomly sampled VQA-DCS probes, three physics-trained annotators assessed four criteria (validity rates and Gwet's AC1 coefficients are shown). For 200 generated images, we report the human-VLM agreement rate and inter-annotator reliability (Gwet's AC1).

Evaluation aspect	Validity / Agree.	Gwet's AC1
VQA-DCS probes (N=50)		
VQA focus & corr.	98.0%	0.896
DCS physical corr.	96.0%	0.892
DCS visual feas.	96.0%	0.892
VQA-DCS sem. cons.	100.0%	0.986
Image labeling (N=200)		
Human-VLM agreement	80.0%	—
Inter-annotator agree.	65.8% (3-way)	0.637

PKPs, including *Refraction of light* and *Diffraction of light*, after instruction refinement for visual displacement and fringe patterns.

5.4 Human Evaluation

To validate the reliability of our automated evaluation pipeline and address concerns about evaluator bias, we conducted a human study with three physics-trained annotators. First, we randomly selected 50 LLM-generated VQA–DCS probe pairs anchored to PKG PKPs and evaluated each on four binary criteria: (i) **VQA focus & correctness**: whether the VQA question targets the core physical phenomenon and its answer is physically correct; (ii) **DCS physical correctness**: whether the descriptive consistency statement conforms to established physical laws; (iii) **DCS visual feasibility**: whether the described state can be unambiguously observed in a static 2D image; and (iv) **VQA–DCS consistency**: whether the VQA question and the DCS description refer to the same underlying physical state. Second, we assessed the alignment between VLM scoring and human judgment on 200 generated images (50 prompts $\times$ 4 images from FLUX.1-dev).

As shown in Table 4, all four probe criteria achieved high validity rates (96%–100%) with substantial to almost perfect inter-annotator agreement (Gwet's AC1: 0.892–0.986). On the 200 generated images, the VLM agreed with human majority at 80.0%, and overall annotator consensus was moderate-to-substantial (AC1=0.637). These results confirm the physical soundness of our benchmark probes and the reliability of VLM-based scoring.

5.5 Robustness and Generalization

We evaluate the stability and generalizability of our framework by analyzing optimization trajectories and cross-model transferability.

Collective Feedback and Stability. As evidenced in our sensitivity study over batch size $B \in \{2, 4, 6, 8, 10, 12\}$, OmniPrompt maintains a robust performance plateau for JanusPro-7B (0.239–0.254), whereas instance-level TextGrad optimization exhibits volatile fluctuations. This stability is primarily attributed to *batch-level* collective feedback: linguistic gradients are derived only after merging B per-query buffers, which aggregates batch-level linguistic gradients to filter stochastic image artifacts and surface shared physical misconceptions. A comparative diagnostic trace of TextGrad versus OmniPrompt gradients is provided in Appendix D.

Cross-Model Policy Generalization. We assess the transferability of evolved policies to unseen closed-source engines. As illustrated in Figure 8, meta-policies developed on open-source backbones yield consistent performance gains across diverse targets. Notably, SeedDream-4.0 achieves a 45.7% relative gain in Joint Score when guided by the Lumina-evolved policy. The fact that policies from various backbones yield uniformly high gains suggests that our optimization captures transferable physical reasoning patterns rather than backbone-specific wording alone. These results support the use of the evolved meta-policy as a physics-aware prompt enhancement across diverse generative paradigms.

6 Conclusion

This paper introduces **OmniPhys**, a benchmark grounded in a PKG to diagnose reasoning deficits in T2I models. Our evaluation reveals systemic bottlenecks in structural and optical consistency, where models prioritize statistical pixel correlations over physical laws. To address these failures, we propose **OmniPrompt**, an iterative framework that distills collective feedback to evolve robust and "physics-aware" meta-policies. Experimental results demonstrate that our approach significantly enhances physical alignment and generalizes across diverse generative backbones.

Acknowledgments

This work is funded by National Natural Science Foundation of China (NSFCU23B2055/NSFC62306276), New Generation Artificial Intelligence-National Science and Technology Major Project 2030 (2025ZD0122800), Yongjiang Talent Introduction Programme (2022A-238-G), and Fundamental Research Funds for the Central Universities (226-2023-00138). This work was supported by Ant Group.

References

[1] Jingkun An, Yinghao Zhu, Zongjian Li, Enshen Zhou, Haoran Feng, Xijie Huang, Bohua Chen, Yemin Shi, and Chengwei Pan. 2025. AGFSync: Leveraging AI-Generated Feedback for Preference Optimization in Text-to-Image Generation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 1746–1754. doi:10.1609/AAAI.V39I2.32168

[2] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, Yimeng Wang, Kai Yu, Wenxuan Chen, Ziwei Feng, Zijian Gong, Jianzhuang Pan, Yi Peng, Rui Tian, Siyu Wang, Bo Zhao, Ting Yao, and Tao Mei. 2025. HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. *CoRR* abs/2505.22705 (2025). arXiv:2505.22705 doi:10.48550/ARXIV.2505.22705

[3] Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2025. R2I-Bench: Benchmarking Reasoning-Driven Text-to-Image Generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, 12595–12630. doi:10.18653/V1/2025.EMNLP-MAIN.636

[4] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *CoRR* abs/2501.17811 (2025). arXiv:2501.17811 doi:10.48550/ARXIV.2501.17811

[5] Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and Wenjia Zhu. 2025. Seedream 4.0: Toward Next-generation Multimodal Image Generation. *CoRR* abs/2509.20427 (2025). arXiv:2509.20427 doi:10.48550/ARXIV.2509.20427

[6] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. 2025. Emu3.5: Native Multimodal Models are World Learners. *CoRR* abs/2510.26583 (2025). arXiv:2510.26583 doi:10.48550/ARXIV.2510.26583

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. https://openreview.net/forum?id=FPnUhsQJ5B

[8] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2023. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models. *CoRR* abs/2305.16381 (2023). arXiv:2305.16381 doi:10.48550/ARXIV.2305.16381

[9] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=PUIqjT4rzq7

[10] Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. 2024. Commonsense-T2I Challenge: Can Text-to-Image Generation Models Understand Commonsense? *CoRR* abs/2406.07546 (2024). arXiv:2406.07546 doi:10.48550/ARXIV.2406.07546

[11] Google DeepMind. 2025. Gemini 3 Pro Image (Nano Banana Pro). https://ai.google.dev/gemini-api/docs/gemini-3. Accessed: 2026-02-09.

[12] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. 2025. Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 15733–15744. doi:10.1109/CVPR52734.2025.01467

[13] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing Prompts for Text-to-Image Generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/d346d91999074dd8d6073d4c3b13733b-Abstract-Conference.html

[14] Nailei Hei, Qianyu Guo, Zihao Wang, Yan Wang, Haofen Wang, and Wenqiang Zhang. 2024. A User-Friendly Framework for Generating Model-Preferred Prompts in Text-to-Image Synthesis. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 2139–2147. doi:10.1609/AAAI.V38I3.27986

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP (1)*. Association for Computational Linguistics, 7514–7528.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NIPS*. 6626–6637.

[17] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract-Datasets_and_Benchmarks.html

[18] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *CoRR* abs/2506.15742 (2025). arXiv:2506.15742 doi:10.48550/ARXIV.2506.15742

[19] Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. PhyBench: A Physical Commonsense Benchmark for Evaluating Text-to-Image Models. *CoRR* abs/2406.11802 (2024). arXiv:2406.11802 doi:10.48550/ARXIV.2406.11802

[20] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. 2025. WISE: A World Knowledge-Informed Semantic Evaluation for Text-to-Image Generation. *CoRR* abs/2503.07265 (2025). arXiv:2503.07265 doi:10.48550/ARXIV.2503.07265

[21] OpenAI. 2024. GPT-4o API Documentation (Snapshot: 2024-11-20). https://platform.openai.com/docs/models/gpt-4o.

[22] J.Z. Pan, G. Vetere, J.M. Gomez-Perez, and H. Wu (Eds.). 2017. *Exploiting Linked Data and Knowledge Graphs for Large Organisations*. Springer.

[23] Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge* (2023), 1–38.

[24] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 7957–7968. doi:10.18653/V1/2023.EMNLP-MAIN.494

[25] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Manyuan Zhang, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Conghui He, Bo Zhang, Xiaohong Liu, Hongsheng Li, Yu Qiao, Chang Xu, and Peng Gao. 2025. Lumina-Image 2.0: A Unified and Efficient Image Generative Framework. *CoRR* abs/2503.21758 (2025). arXiv:2503.21758 doi:10.48550/ARXIV.2503.21758

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. doi:10.1109/CVPR52688.2022.01042

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html

[28] Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *CoRR* abs/2507.06261 (2025). arXiv:2507.06261 doi:10.48550/ARXIV.2507.06261

[29] Z.-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven C. H. Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. 2025. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer. *CoRR* abs/2511.22699 (2025). arXiv:2511.22699 doi:10.48550/ARXIV.2511.22699

[30] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente

Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *CoRR* abs/2503.20314 (2025). arXiv:2503.20314 doi:10.48550/ARXIV.2503.20314

[31] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025. Qwen-Image Technical Report. *CoRR* abs/2508.02324 (2025). arXiv:2508.02324 doi:10.48550/ARXIV.2508.02324

[32] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html

[33] Cilin Yan, Jingyun Wang, Lin Zhang, Ruihui Zhao, Xiaopu Wu, Kai Xiong, Qingsong Liu, Guoliang Kang, and Yangyang Kang. 2025. Efficient and Accurate Prompt Optimization: the Benefit of Memory in Exemplar-Guided Reflection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 753–779. https://aclanthology.org/2025.acl-long.37/

[34] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large Language Models as Optimizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=Bb4VGOWELI

[35] Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. 2025. WorldGenBench: A World-Knowledge-Integrated Benchmark for Reasoning-Driven Text-to-Image Generation. *CoRR* abs/2505.01490 (2025). arXiv:2505.01490 doi:10.48550/ARXIV.2505.01490

A Implementation of System Prompt Initialization

System Prompt Specification for $\mathcal{L}$

Role: You are a prompt expander for image generation. Given a simple input prompt, your job is to expand it into a rich, physically plausible, visually descriptive sentence that guides image models to generate accurate and realistic scenes.

Step 1: Physical Reasoning
Reason about the implicit physical states of objects in the prompt. Think through what physical principles are involved and how they influence the object's behavior.

Step 2: Generate the expanded descriptive text.
Combine the original prompt with the physical insights from Step 1 to produce a rich, coherent, and visually specific description—optimized for image generation.

Finally, output your expanded text in JSON format: {"expanded_text": "The final expanded text for image generation — descriptive, concrete, and physics-aware."}

B Stability Analysis of Collective Linguistic Gradients

As illustrated in Figure 9, the scale of collective feedback B plays a critical role in balancing the precision of linguistic gradients with computational efficiency. In Figure 9 (a), we observe that JanusPro-7B is highly robust to variations in collective scale, suggesting a relatively smooth optimization landscape. In contrast, FLUX.1-dev exhibits a sharp performance drop outside the $B \in [6, 10]$ range, indicating that large-scale Diffusion Transformers (DiTs) are more susceptible to the density of meta-level linguistic gradients. Furthermore, Figure 9 (b) highlights the anti-noise capability inherent in our collective feedback mechanism; while instance-level methods (e.g., TextGrad) suffer from stochastic oscillations triggered by outlier images in the T2I manifold, `OmniPrompt` leverages cross-sample consensus to maintain a steady upward trajectory toward physical alignment.

C Illustration of PKG-Guided Coupled Probe Generation

For each implicit query x, OmniPhys retrieves the anchor PKP's atomic physical statements and prompts an LLM with a visual-physical template, and PKP-aligned few-shot demonstrations. The LLM selects n scene-specific physical state points and co-generates coupled units $u_i = (\text{VQA}_i, \text{DCS}_i)_{i=1}^{n}$, where DCS denotes a Descriptive Consistency Statement (Section 3.2).

Case: Reflection in a Flat Mirror (*Reflection of light, $n = 2$*)

Input. x: "`A cylindrical candle sits on a table facing a wall-mounted mirror.`" Atomic statements: equal incidence/reflection angles; object–image equidistance to the mirror.

Generated units.

- **Unit 1 (size):** VQA: "Is the reflected candle smaller than the actual candle?" **Ans:** no; DCS: same height and width in object and reflection.
- **Unit 2 (distance):** VQA: "Is the reflection closer to the mirror than the candle?" **Ans:** no; DCS: reflection is equidistant behind the mirror; candle-to-reflection distance is twice the object-to-mirror distance.

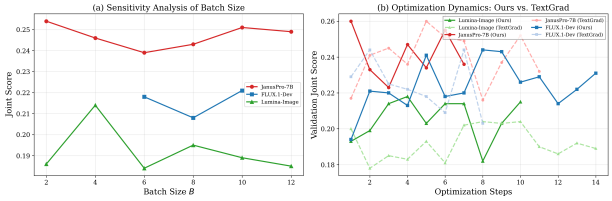

Figure 9: Sensitivity analysis and optimization dynamics. (a) Joint Scores across different batch sizes B. (b) Step-wise validation trajectories comparing our collective optimization against TextGrad.

D Diagnostic Trace of Evolved Linguistic Gradients

Table 5 contrasts the gradient-update behavior of **TextGrad** versus **OmniPrompt**. Both runs share the same initial meta-policy $P^{(0)}$,

Table 5: TextGrad vs. OmniPrompt: representative linguistic-gradient traces (excerpts abbreviated from our logs).

TextGrad (single-query backward)	**OmniPrompt ($B = 2$ batch merge)**
Stage 1: $g_{\tilde{x}}$	**Stage 1: unified** $g_{\tilde{x}}$
Excerpt: "The primary issue is a dense plastic bag floating in mid-air, violating gravity and buoyancy… suggest tethering, external support, or a lighter-than-air fill."	*Excerpt:* One merged critique spans the batch: (i) "inflatable toy floating in mid-air without visible support or buoyancy mechanism"; (ii) vague melt/heat wording that should specify concrete transformation (e.g., softening, bubbling).
Analysis: Feedback is tied to **one query's** dominant failure at a time.	*Analysis:* Batch merge produces **one** $g_{\tilde{x}}$ covering multiple failure modes across the two queries.
Stage 2: g_P	**Stage 2:** $\{g_P^{(i)}\}_{i=1}^{2}$
Excerpt: "Add explicit emphasis on material and physical properties… describe material state (gas/liquid/solid) and container interaction."	*Excerpt:* "Strengthen Reason-before-Synthesis… ensure physical laws hold and **state the mechanism** (e.g., buoyancy, suspension)." (two branch critiques aggregated by Φ.)
Analysis: Meta-gradient mainly adds **content-level** directives.	*Analysis:* Meta-gradient targets the **reasoning procedure** in P.
Stage 3: $P^{(t+1)}$	**Stage 3:** $P^{(t+1)}$
Excerpt: 5 steps: analyze states → environment → generate → verify → refine iteratively.	*Excerpt:* 8 steps: physical principles → material → environment → cause–effect → visual cues → edge cases → review → JSON.
Analysis: Policy grows via appended rules after query-local failures.	*Analysis:* Policy becomes longer but more **explicitly staged** (principles, materials, validation).

evaluator $\mathcal{E}$, and ensemble size $K = 4$, and include the same gas-density queries (e.g., a helium-filled inflatable toy in CO_2 and an SF_6-filled plastic bag in H_2).

TextGrad performs *single-query* backward: each query is evaluated and backpropagated separately, producing query-local $g_{\tilde{x}}$ and g_P before Φ rewrites P. **OmniPrompt** performs *batch-merged* backward with $B = 2$: forward passes accumulate two queries, merged $(\tilde{x}_i, e_i)$ yield one unified $g_{\tilde{x}}$, two branch-specific meta-gradients $\{g_P^{(i)}\}_{i=1}^{2}$, and a single Φ update. Stages 1–3 below denote $g_{\tilde{x}}$, meta-gradients, and $P^{(t+1)}$ (Eqs. 4–6).

E End-to-End Illustration of OmniPrompt

This appendix walks through **once OmniPrompt update** with $B = 2$ and $K = 4$, complementing Section 4. External modules: $\mathcal{G}$ (JanusPro-7B), $\mathcal{E}$ (Gemini-2.5-Pro); graph participants: $\mathcal{L}$, P, Γ, Φ (GPT-4o). Full instruction templates and the complete training log are in our code release.

E.1 Running Example ($B = 2$, $K = 4$)

Stage	**Module**	**Output**
I	$\mathcal{L}$	$i = 1$: x_1 (helium toy in sealed CO_2 room) → $\tilde{x}_1$. $i = 2$: x_2 (SF_6 bag in sealed H_2 room) → $\tilde{x}_2$.
II	$\mathcal{G}, \mathcal{E}$	For each $i \in \{1, 2\}$: sample $K = 4$ images and form $e_i = \langle \bar{S}_i, \mathcal{R}_i \rangle$ (e.g., $\bar{S}_1 = 0.50$ with unsupported-hovering violations).
III	loss node	Batch merge: $\bigoplus_{i=1}^{2} \tilde{x}_i$ and $\bigoplus_{i=1}^{2} e_i$ → TextGrad root.
IV	Γ	One unified $g_{\tilde{x}}$; two meta-gradients $g_P^{(1)}, g_P^{(2)}$ (one per query branch).
V	Φ	Aggregate $\{g_P^{(1)}, g_P^{(2)}\}$ → single update $P^{(t+1)}$.

Fixed templates (abbreviated). **Meta-policy** $P^{(0)}$**.** Reasoning-before-Synthesis: infer physical states, then emit JSON `{"expanded_text": ...}`. **Evaluator** $\mathcal{E}$ **(training).** VLM compares each image against the query's *pre-generated DCS probes*; output format: `Score: [0|1]. Reason: [one sentence]`. Per-image scores are aggregated into $e_i = \langle \bar{S}_i, \mathcal{R}_i \rangle$ (Eq. 3). Γ **and** Φ**.** Standard TextGrad string-function backward and Textual Gradient Descent; batch loss binds $\bigoplus_i \tilde{x}_i$ to merged feedback, yielding one $g_{\tilde{x}}$ and $\{g_P^{(i)}\}_{i=1}^{B}$, then one rewrite of P.