# FeynmanBench: Benchmarking Multimodal LLMs on Diagrammatic Physics Reasoning

**Anonymous ACL submission**

## Abstract

Current multimodal benchmarks for scientific reasoning primarily evaluate local information extraction—models recognize symbols and values and then perform textual inference. They do not assess whether models can reason over the global structural properties of formal diagrams, such as topology, conservation constraints, and the consistent mapping between visual patterns and algebraic expressions. We introduce FeynmanBench, a benchmark of over 2,000 tasks centered on Feynman diagrams spanning the electromagnetic, weak, and strong interactions of the Standard Model. Each instance couples a diagram image with minimal textual conventions and requires models to recover the full physical content—vertex inventory, propagator types, topological connectivity, momentum routing, and the complete scattering amplitude. An automated generation and verification pipeline produces the diagrams, annotations, and reference answers under standardized rules. Evaluating 19 state-of-the-art multimodal LLMs, we find a consistent failure pattern: models achieve 70–95% on local recognition (vertex and propagator identification) but collapse to 13–17% on topological reconstruction (CP3), and near zero on full algebraic derivation (CP5). FeynmanBench offers a controlled testbed for multimodal reasoning over formal scientific diagrams and highlights fundamental limitations of current architectures in topology-sensitive scientific reasoning.

**Keywords:** Multimodal LLMs, Feynman Diagrams, Visual Reasoning, Benchmark

## 1 Introduction

Diagrammatic reasoning is a core methodology in modern physics and mathematics (Olivier, 2001). Among diagrammatic formalisms, Feynman diagrams (Feynman, 1949; Kumerički, 2016) play a unique role: they encode perturbative expansions in quantum field theory as graph structures, mapping intuitive visual representations to rigorous computation via unambiguous Feynman rules. Each diagram is a graph whose connectivity, conservation laws, and algebraic derivation must satisfy precise global constraints. Feynman diagrams are especially suitable for benchmarking because they combine three properties: a rigid visual syntax (vertices, propagators, and external legs follow fixed graphical conventions), global physical constraints (conservation laws, gauge symmetries, and the topological identity $L = I - V + 1$), and machine-verifiable symbolic outputs (amplitudes can be generated and checked automatically). Unlike many scientific images that admit descriptive interpretation, a Feynman diagram has tightly constrained semantics where small topological differences produce different physical processes and amplitudes. This makes them an ideal testbed for evaluating whether multimodal models can perform structured visual reasoning rather than superficial pattern matching.

Existing multimodal benchmarks in the physical sciences (He et al., 2024; Wang et al., 2025b; Feng et al., 2025; Chen et al., 2023) predominantly follow a "recognition then reasoning" paradigm: models first extract discrete symbols via OCR or object recognition, then perform textual derivation using internalized formulas. While useful for measuring parameter extraction and textbook problem-solving, this paradigm does not test whether models can reason directly over the global structure of a diagram. Topological relationships, cross-regional constraints, and implicit physical symmetries—the defining features of formal scientific notation—remain unexamined (Xu et al., 2025; Al-Tahan et al., 2024). Recent benchmarks such as PhyX (Shen et al., 2025) and MAPS (Zhu et al., 2025) reduce textual dependence or introduce physics engines, but still either generate intermediate descriptions or shift evaluation toward tool-calling accuracy.

We introduce FeynmanBench, a benchmark de-

signed to evaluate multimodal LLMs on tasks that require simultaneous diagrammatic recognition, topological reasoning, and formal algebraic derivation directly in the visual domain. The benchmark comprises over 2,000 Feynman diagrams drawn from more than 100 distinct topological types, covering electromagnetic, electroweak, and strong interactions. Each instance provides a diagram image together with minimal textual conventions (interaction type, external leg identities, and momentum orientation)—sufficient only to resolve drawing-convention ambiguities, not to supply topological or algebraic answers. Models must report the vertex inventory (CP1), propagator content (CP2), topological connectivity (CP3), momentum routing (CP4), and the full scattering amplitude with symmetry factors (CP5). An automated pipeline based on FeynArts and FeynCalc (Hahn, 2001; Shtabovenko et al., 2016) generates the diagrams, topological annotations, and reference amplitudes under standardized gauge conventions, ensuring scalability and verifiability.

In summary, our core contributions are: (1) FeynmanBench, the first benchmark for evaluating multimodal diagrammatic reasoning over Feynman diagrams; (2) an automated generation and verification pipeline that produces diagrams, annotations, and reference amplitudes under standardized conventions; and (3) a systematic evaluation of 19 MLLMs exposing the gap between local visual recognition and global structural reasoning. Our experiments reveal a consistent bottleneck: local symbol recognition does not translate into global structural understanding. The top-scoring model achieves 52.5% overall, while the strongest full-benchmark models plateau around 28–30%. Across all models, we observe a consistent collapse at CP3 (topological connectivity)—even models with 94.7% vertex accuracy drop to 16.8% on topology—and CP5 (algebraic/symmetry factors) remains below 2% for most. Strikingly, we find that reducible topologies are consistently harder than irreducible ones, suggesting that models rely on memorized textbook patterns rather than transferable structural deduction. Our code, pipeline, and benchmark will be released upon acceptance.

## 2 Related Work

**Multimodal Scientific Benchmarks.** Benchmarks for multimodal scientific reasoning have progressed from undergraduate textbooks to Olympiad-level problems (He et al., 2024; Feng et al., 2025). TheoremQA (Chen et al., 2023) and PhysUniBench (Wang et al., 2025b) require models to extract numerical parameters from images and apply formulas, while MM-PhyQA (Anand et al., 2024) uses multi-image chain-of-thought prompting. A consistent limitation unites these efforts: evaluation focuses on the combined accuracy of parameter extraction and textual logic, leaving the visual reasoning component unmeasured. PhyX (Shen et al., 2025) reduces textual cues to isolate visual dependence, but still allows models to generate structured descriptions as an intermediate step. MAPS (Zhu et al., 2025) and LLM-Phy (Cherian et al., 2024) integrate physical simulators, shifting evaluation toward tool use rather than intrinsic visual reasoning. In parallel, recent visual multimodal models have increasingly unified understanding and generation within single architectures (Zhao et al., 2025; Deng et al., 2025; Wu et al., 2025a), yet systematic evaluation of visual reasoning over formal scientific notation remains under-explored. In contrast, FeynmanBench requires models to reason directly over diagrammatic structure, where correctness is determined by global topology and algebraic consistency rather than local feature extraction. We designed the benchmark so that even a model with perfect OCR capability cannot succeed without genuine visual-structural understanding.

**ML for Feynman Diagrams.** Prior work applies machine learning to accelerate Feynman-diagram computation—predicting scattering matrix elements via GNNs (Mitchell et al., 2022), solving Feynman integrals with PINNs (Calisto et al., 2024), and sampling diagrams via normalizing flows (Leoni and Franchini, 2024). These efforts use ML as a computational tool for physics, but do not benchmark multimodal reasoning over the diagrams themselves. Our work is orthogonal: rather than using ML to solve physics problems, we use physics problems as a probe for ML reasoning capability.

## 3 FeynmanBench

### 3.1 Task Definition

Each instance in FeynmanBench consists of two inputs: (1) a Feynman diagram image, and (2) minimal textual conventions specifying the interaction type (e.g., QED, weak, QCD), the identities and momentum labels of external legs, and the dia-

gram orientation. These conventions are intentionally limited—they resolve drawing ambiguity (e.g., momentum-direction conventions for antiparticles) without revealing any topological or algebraic answer. The model must produce five outputs, evaluated via checkpoints CP1–CP5:

**CP1 – Vertex inventory:** count and type of each interaction vertex (e.g., $\bar{\psi}\gamma^\mu\psi A_\mu$, $W^+W^-Z$), with the correct linked field count per vertex.

**CP2 – Propagator inventory:** complete list of internal lines with correct particle types and multiplicities.

**CP3 – Topological connectivity:** graph of vertex connections isomorphic to the original diagram.

**CP4 – Momentum routing:** self-consistent internal momentum assignment satisfying conservation at each vertex.

**CP5 – Full amplitude:** algebraic expression including Dirac structures, trace contractions, Lorentz index closures, global signs, fermion-loop factors, and combinatorial symmetry factors.

The five checkpoints form a deliberate progression. CP1 and CP2 test whether the model can identify primitives; CP3 and CP4 test whether it can assemble those primitives into a self-consistent physical graph; CP5 tests whether that graph-level understanding is sufficient to support full symbolic derivation. Each checkpoint is evaluated independently and unconditionally (a model need not pass earlier checkpoints to be scored on later ones), but in practice errors cascade: an incorrect vertex in CP1 invalidates the topology in CP3, which in turn makes the amplitude in CP5 unsalvageable. The overall model score is the arithmetic mean of the five unconditional checkpoint pass rates.

These conventions are necessary because certain drawing choices in professional Feynman diagrams are not semantically unique in isolation—momentum-arrow conventions for antiparticles, omitted charge labels for some bosons, and similar notational ambiguities arise in multi-purpose diagrams. The text therefore disambiguates notation without supplying structural or algebraic answers: the critical reasoning for CP3–CP5 remains entirely visual-structural.

### 3.2 Dataset Generation

We construct FeynmanBench using an automated pipeline built on FeynArts and FeynCalc (Hahn, 2001; Shtabovenko et al., 2016). We first enumerate connected graph topologies (vertex-line skeletons) with specified numbers of external legs and loops. We then insert physical fields by mapping particle content from Standard Model definition files onto the skeleton lines, enforcing vertex-level conservation laws and coupling rules. Finally, we convert field-inserted diagrams to analytic amplitudes by applying symbolic Feynman rules—propagators for internal lines, vertex factors for interaction points, and combinatorial symmetry factors from topological automorphisms. All amplitudes are generated in Feynman gauge.

For each diagram, we store: (1) the rendered image; (2) external-leg metadata; (3) vertex-level type and connectivity annotations; (4) internal propagator annotations; (5) a canonicalized topology record; and (6) the reference amplitude expression. This structured representation supports automated evaluation across all checkpoints. From the full candidate pool of over 200,000 diagrams, we select the benchmark subset via stratified sampling over topology class (A1–A5) and interaction sector (B1–B6), removing near-duplicate layouts to maximize structural diversity under a fixed evaluation budget.

### 3.3 Data Taxonomy

Diagrams are classified along two dimensions:

- **Topology complexity (A1–A5):** A1: tree-level; A2: one-loop one-particle-reducible (1PR); A3: one-loop one-particle-irreducible (1PI); A4: two-loop 1PR; A5: two-loop (1PI). A diagram is reducible if cutting a single internal line disconnects the graph. Reducible topologies are less common in textbooks and thus test whether models rely on memorized patterns or perform genuine structural deduction.

- **Interaction sector (B1–B6):** B1: QED with photons and $e^\pm$ only; B2: QED with generic charged fermions; B3: electroweak with leptons and neutrinos; B4: electroweak including $W/Z$ gauge bosons; B5: electroweak including the Higgs boson; B6: strong (QCD) with quarks and gluons.

In total, we identify 102 distinct task types, produce over 200,000 candidate diagrams, and select a representative subset of 2,000 that preserves diversity across topology and interaction classes. Table 1 reports the per-category counts. The benchmark is intentionally unbalanced across categories—certain higher-loop electroweak sectors naturally contain many more valid diagrams

than simple QED trees—and we preserve this natural variation while ensuring that all topology and interaction classes remain represented.

## 4 Experiments

### 4.1 Setup

Figure 2 and Table 2 should be read primarily as evidence of a shared failure pattern rather than a leaderboard: across diverse model families and scales, local recognition remains substantially stronger than topology-sensitive reasoning.

**Models.** We evaluate 19 multimodal LLMs from five major providers, spanning the GPT, Gemini, Claude, Qwen, GLM, Kimi, and Doubao families. Due to API rate limits and cost constraints, we evaluate four models (GPT-5.5, Kimi K2.6, Gemini 3.1 Pro, Gemini 3 Flash) on a representative subset ($N$ ranging from 165 to 742), stratified by topology and interaction classes to match the full benchmark distribution; the remaining 15 models are evaluated on the full benchmark ($N \approx 2{,}000$). All models are accessed via API with temperature set to zero. We perform five independent evaluation runs per instance and report the mean pass rate across runs. Although decoding is configured deterministically (temperature = 0), we retain repeated runs because closed-source API systems can exhibit mild nondeterminism due to backend routing and vision preprocessing pipelines; averaging across runs reduces such variance.

**Prompt.** As shown in Fig. 3, we design the prompt to provide only the minimum conventions needed to make the task well-posed: interaction type, external-leg identities, momentum labels, and diagram orientation. These elements disambiguate drawing conventions (e.g., momentum-direction rules for antiparticles) without supplying any topological or algebraic answer. We intentionally provide only these metadata-level cues—the critical reasoning for CP3–CP5 must still be performed over the diagram structure itself.

**Evaluation protocol.** We evaluate the five checkpoints independently. For CP1 and CP2, we use automated rule-based scripts that verify correctness against ground-truth annotations, distinguishing three grades: correct, partially correct (right count but wrong type or linkage), and incorrect. For CP3–CP5, we adopt an LLM-as-a-judge framework with a fixed evaluation prompt, a panel of three independent models, and majority vote to determine the final score. To validate this protocol, we conduct manual spot-checks on 100 randomly sampled instances across CP3–CP5; we find 91% agreement between human judgment and the majority-vote outcome, confirming the reliability of automated scoring.

### 4.2 Results Analysis

We present the full evaluation results in Fig. 2 and Table 2. Our analysis reveals three key findings.

**The CP3 cliff.** We observe a sharp and universal performance collapse from local recognition to topological reconstruction in every model tested. Gemini 3 Flash exemplifies the pattern: 94.7% on CP1 and 84.5% on CP2, dropping to 16.8% on CP3 and 5.0% on CP4. Doubao Seed 2.0 Pro scores 79.1% on CP1 but only 13.3% on CP3. We find this pattern across all model families and scales. This CP3 collapse is the central diagnostic of FeynmanBench: a model can achieve high CP1/CP2 rates by recognizing local symbols and line styles, but CP3 requires constructing a globally coherent latent graph and verifying that distant regions of the image are mutually compatible. The collapse therefore indicates a failure of structural composition rather than a mere loss of low-level visual accuracy. Because four frontier models are evaluated on stratified subsets rather than the full benchmark, we focus our conclusions on performance trends across checkpoints rather than fine-grained leaderboard comparisons.

**Topology complexity and memorization.** We stratify CP4 pass rates by topology class (Table 3) and find sharp degradation with increasing complexity: even top models achieve 40–57% on tree-level diagrams (A1) but collapse to near zero on two-loop structures (A4, A5). Critically, we observe that reducible graphs (A2) are consistently harder than irreducible ones (A3)—Claude Sonnet 4.6 scores 16.2% vs. 28.2%; Qwen 3.6 Plus scores 11.0% vs. 21.4%. This pattern is counter-intuitive from a graph-theoretic perspective, since 1PR graphs are often decomposable into simpler substructures. Their greater difficulty for MLLMs suggests that rarity in training data outweighs compositional simplicity—models rely on memorized patterns from textbooks rather than transferable structural reasoning. We also stratify by interaction type: strong interactions (B6) yield the highest CP4 rates (23.9% for Claude Sonnet 4.6, 24.2% for Qwen 3.6 Plus), while electroweak processes with leptons and neutrinos (B3) are consistently the hardest across all models, reflecting the added com-

Table 1: Diagram counts by interaction sector and topology class. Topology: tree (A1), one-loop 1PR (A2), one-loop 1PI (A3), two-loop 1PR (A4), two-loop 1PI (A5).

| Interaction sector | External-leg content | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|---|
| QED | Photons, $e^{\pm}$ only | 15 | 86 | 17 | 150 | 119 |
| QED | Photons + generic charged fermions | 11 | 112 | 12 | 106 | 72 |
| Electroweak | Leptons and neutrinos | 45 | 125 | 51 | 50 | 120 |
| Electroweak | Includes $W/Z$ | 8 | 115 | 68 | 75 | 147 |
| Electroweak | Includes Higgs | 17 | 56 | 122 | 70 | 89 |
| QCD | Quarks / gluons | 7 | 38 | 74 | 51 | 28 |

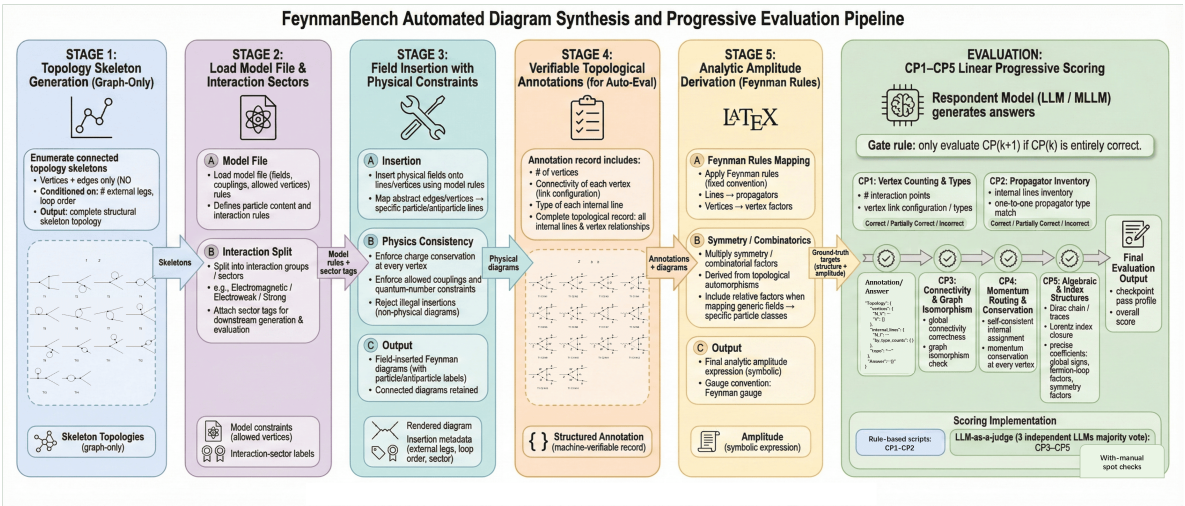


Figure 1: Workflow of FeynmanBench. Panels 1–4: diagram generation and annotation. Panel 5: amplitude derivation via Feynman rules. Panel 6: automated scoring against ground truth. The same pipeline that generates diagrams also produces their verifiable structural and symbolic ground truth, enabling scalable evaluation without manual answer authoring.

plexity of flavor mixing and chiral structures. The heatmaps in Fig. 5 provide a finer-grained view of CP1–CP3 performance across individual topology and interaction categories, confirming that model strengths in local vertex recognition largely disappear when spatial complexity increases.

**CP5: near-zero algebraic reasoning.** We find that no model achieves meaningful CP5 performance—the best (GPT-5.5) reaches only 15.8%, and all full-benchmark models score below 2%. CP5 requires the complete derivation of symmetry factors, Dirac matrix orderings, and overall normalization. Examining the rare CP5 successes, we note that they correspond almost exclusively to diagrams found in published textbooks, further supporting our memorization hypothesis. CP5 is especially demanding because errors compound from earlier stages: an incorrect propagator type from CP2, a missing vertex connection from CP3, or an inconsistent momentum route from CP4 each independently invalidate the final amplitude, even if the symbolic form appears superficially plausible. The near-zero CP5 rates therefore reflect not only symbolic weakness but also the absence of robust graph-level internal representations upstream.

### 4.3 Error Analysis

Through qualitative analysis of model outputs, we identify four recurring failure modes. These errors are not independent: in many cases, an early visual parsing mistake triggers a cascade—a spurious vertex changes the inferred propagator set, which in turn breaks graph connectivity, invalidates momentum conservation, and makes the final amplitude

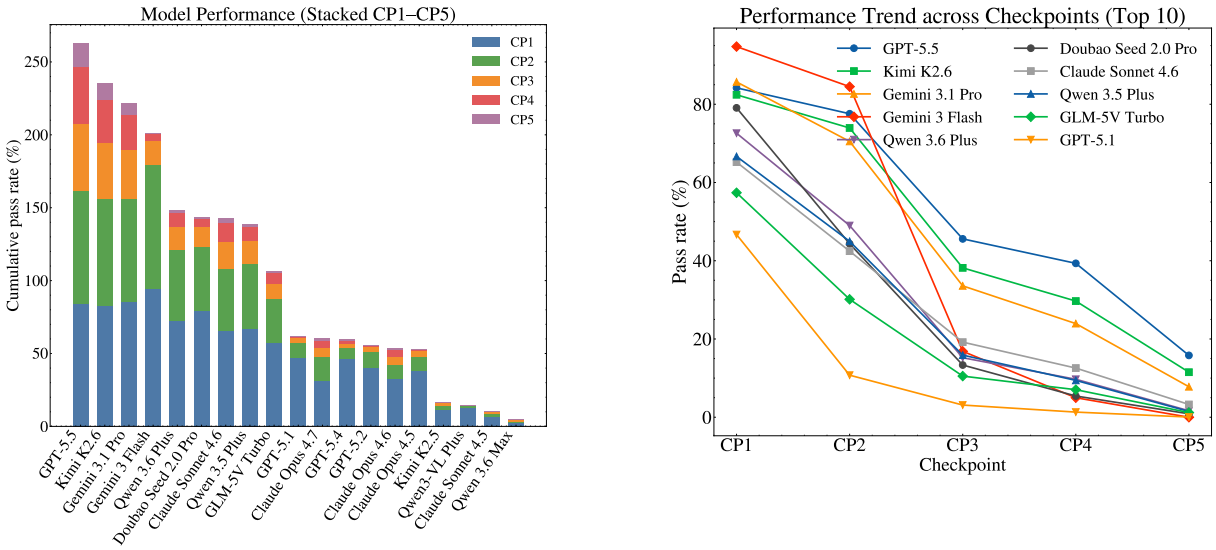


Figure 2: Model performance across checkpoints.

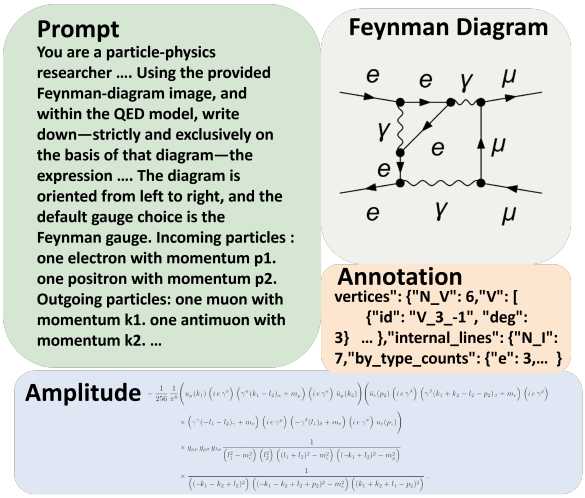


Figure 3: Prompt structure and task definitions with example expectations. The prompt provides minimal conventions (interaction type, external-leg identities, momentum orientation)—sufficient to resolve drawing ambiguities without supplying topological or algebraic answers.

unsalvageable. This cascading pattern helps explain why CP5 remains near zero even when CP1 and CP2 scores are moderately strong. Figure 4 illustrates representative examples:

- **Visual parsing failures.** We find that models frequently mistake incidental line crossings in 2D projections for interaction vertices, inflating vertex counts and corrupting downstream topological reconstruction.

- **Vision-logic decoupling.** When inferring internal-line identities, models fail to apply conservation laws (charge, flavor, baryon number). We consistently observe that charged bosons ($W^{\pm}$, $G^{\pm}$) are conflated with neutral counterparts ($G^0$, $Z$), even though the external-leg charge configuration determines the internal identity uniquely.

- **Graph-theoretic inconsistency.** The topological identity $L = I - V + 1$ (relating loops $L$, internal lines $I$, and vertices $V$) is routinely violated in model outputs. We note that models report $V$, $I$, and $L$ values that cannot simultaneously satisfy this basic graph-theoretic constraint, indicating a fundamental lack of self-consistency checking.

- **Ignoring symmetry and chirality.** We observe that models make almost no attempt to compute combinatorial symmetry factors or to invoke chiral projection operators ($V - A$ structures), even when explicitly prompted. We believe determining a symmetry factor requires identifying the automorphism group of the graph—a task demanding precise global structural mastery that current architectures lack. This failure mode directly explains why CP5 remains near zero: even a correctly reconstructed topology cannot yield a correct amplitude without the proper symmetry factor.

## Limitations

We acknowledge several limitations of the current work. First, FeynmanBench covers Standard

Table 2: Overall checkpoint pass rates (%). Score $= \frac{1}{5}\sum_j \mathrm{CP}_j$. †: evaluated on a stratified subset. ††: not covered for this category.

| Model | $N$ | CP1 | CP2 | CP3 | CP4 | CP5 | Score |
|---|---|---|---|---|---|---|---|
| GPT-5.5† | 272 | 84.2 | 77.6 | 45.6 | 39.3 | 15.8 | 52.5 |
| Kimi K2.6† | 165 | 82.4 | 73.9 | 38.2 | 29.7 | 11.5 | 47.2 |
| Gemini 3.1 Pro† | 539 | 85.7 | 70.5 | 33.6 | 23.9 | 7.8 | 44.3 |
| Gemini 3 Flash† | 742 | 94.7 | 84.5 | 16.8 | 5.0 | 0.0 | 40.2 |
| Qwen 3.6 Plus | 2051 | 72.6 | 49.0 | 15.2 | 9.7 | 1.7 | 29.6 |
| Doubao Seed 2.0 Pro | 2046 | 79.1 | 44.3 | 13.3 | 5.4 | 0.9 | 28.6 |
| Claude Sonnet 4.6 | 2021 | 65.2 | 42.5 | 19.2 | 12.6 | 3.3 | 28.5 |
| Qwen 3.5 Plus | 2028 | 66.7 | 45.0 | 15.9 | 9.4 | 1.5 | 27.7 |
| GLM-5V Turbo | 1931 | 57.4 | 30.1 | 10.5 | 7.0 | 1.2 | 21.3 |
| GPT-5.1 | 2056 | 46.7 | 10.7 | 3.1 | 1.3 | 0.0 | 12.4 |
| Claude Opus 4.7 | 2052 | 31.4 | 16.6 | 5.9 | 4.7 | 1.5 | 12.0 |
| GPT-5.4 | 2056 | 46.3 | 7.7 | 3.0 | 2.1 | 0.9 | 12.0 |
| GPT-5.2 | 2056 | 39.9 | 11.2 | 3.3 | 0.9 | 0.0 | 11.1 |
| Claude Opus 4.6 | 2046 | 32.6 | 9.4 | 5.7 | 4.5 | 1.0 | 10.7 |
| Claude Opus 4.5 | 2052 | 38.6 | 9.6 | 3.5 | 0.9 | 0.0 | 10.5 |
| Kimi K2.5 | 1871 | 11.2 | 3.4 | 1.4 | 0.2 | 0.0 | 3.2 |
| Qwen3-VL Plus | 2044 | 12.3 | 2.0 | 0.3 | 0.0 | 0.0 | 2.9 |
| Claude Sonnet 4.5 | 2035 | 6.5 | 2.0 | 1.1 | 0.6 | 0.0 | 2.0 |
| Qwen 3.6 Max | 2053 | 2.5 | 1.0 | 0.8 | 0.5 | 0.2 | 1.0 |

Table 3: CP4 pass rates (%) by topology class (A1–A5), isolating the effect of graph complexity. Without this stratification, aggregate scores would obscure the fact that most models fail almost completely once reasoning requires multi-loop global consistency. †: subset. "–" denotes category not covered.

| Model | A1 | A2 | A3 | A4 | A5 | Overall |
|---|---|---|---|---|---|---|
| Kimi K2.5† | 75.0 | 25.0 | – | – | – | 52.3 |
| GPT-5.5† | 57.6 | 28.9 | – | – | – | 39.3 |
| Kimi K2.6† | 53.3 | 20.8 | – | – | – | 29.7 |
| Gemini 3.1 Pro† | 39.8 | 18.0 | 32.2 | 8.2 | 27.6 | 23.9 |
| Claude Sonnet 4.6 | 56.9 | 16.2 | 28.2 | 0.2 | 3.7 | 12.6 |
| Qwen 3.6 Plus | 48.5 | 11.0 | 21.4 | 0.6 | 3.3 | 9.7 |
| Qwen 3.5 Plus | 47.5 | 10.5 | 21.3 | 0.2 | 3.4 | 9.4 |
| GLM-5V Turbo | 42.9 | 8.5 | 15.0 | 0.2 | 0.7 | 7.0 |
| Doubao Seed 2.0 Pro | 48.5 | 5.8 | 8.7 | 0.0 | 0.3 | 5.4 |
| Claude Opus 4.7 | 45.6 | 5.2 | 6.7 | 0.0 | 0.0 | 4.7 |
| Claude Opus 4.6 | 44.0 | 1.9 | 11.7 | 0.0 | 0.5 | 4.5 |
| GPT-5.4 | 26.2 | 2.1 | 1.0 | 0.0 | 0.3 | 2.1 |

*Note:* Each rate is #(CP4 correct)/#(total items in category).

Model perturbative diagrams up to two-loop order under Feynman gauge; it does not include gravitational interactions, non-perturbative regimes, or higher-loop contributions. Second, we evaluated four frontier models on stratified subsets ($N = 165$–$742$) rather than the full benchmark; while the subset preserves category-level distributions, these smaller samples reduce statistical resolution for subgroup analyses. We also note that our LLM-as-a-judge protocol for CP3–CP5 was spot-checked on 100 instances (91% agreement with majority vote) but has not been independently verified across every instance, and scoring sensitivity to prompt variations remains unexamined. Third, all diagrams are machine-generated under standardized rendering; model behavior on hand-drawn or photographed diagrams with visual noise may differ. Finally, our benchmark evaluates single-pass diagram comprehension under fixed textual conventions, and our model coverage is limited to proprietary API-accessible systems without open-weight models or human expert baselines for calibration.

## 5 Conclusion

We introduced FeynmanBench, a benchmark of over 2,000 Feynman-diagram tasks with automated generation and verification, designed to evaluate multimodal LLMs on diagrammatic reasoning that requires global structural understanding. Evaluating 19 models reveals a consistent failure mode: local visual recognition is relatively strong, but performance collapses at topological reconstruction (CP3) and is virtually absent at full algebraic derivation (CP5). The inversion of reducible vs. irreducible difficulty further suggests reliance on memorization over structural deduction. These results indicate that the primary bottleneck for formal scientific diagrams is not visual perception but the maintenance of globally consistent representations under strict constraints. We believe this gap between local parsing and global reasoning will persist for any notation system whose semantics are defined relationally rather than compositionally.

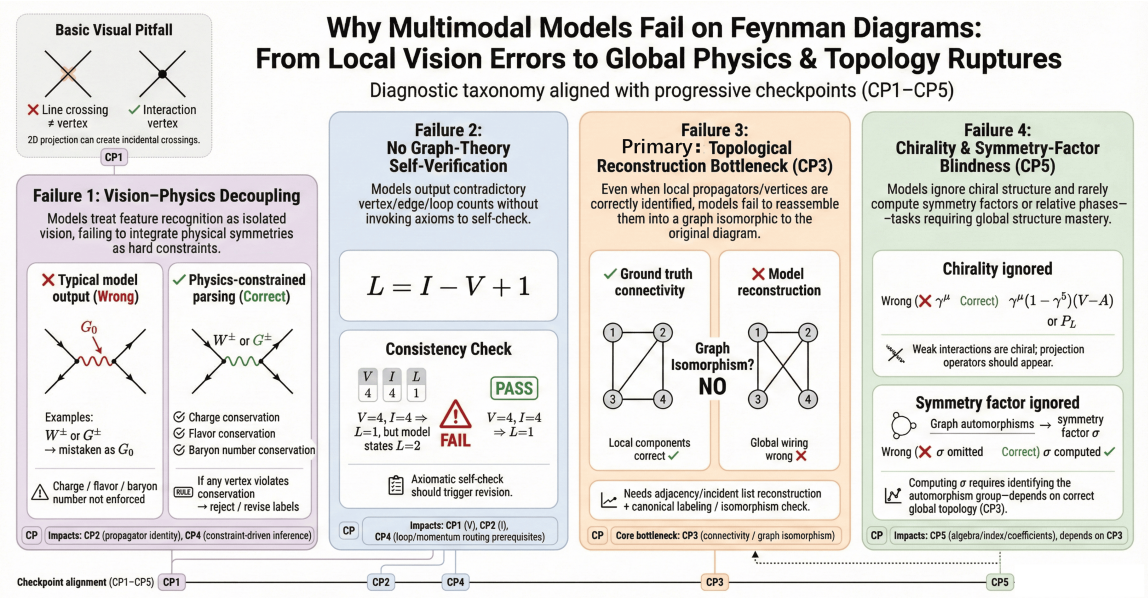


Figure 4: Representative examples of the four dominant failure modes. Each case corresponds to a pattern repeatedly observed across models and interaction sectors, rather than an isolated anomalous output.

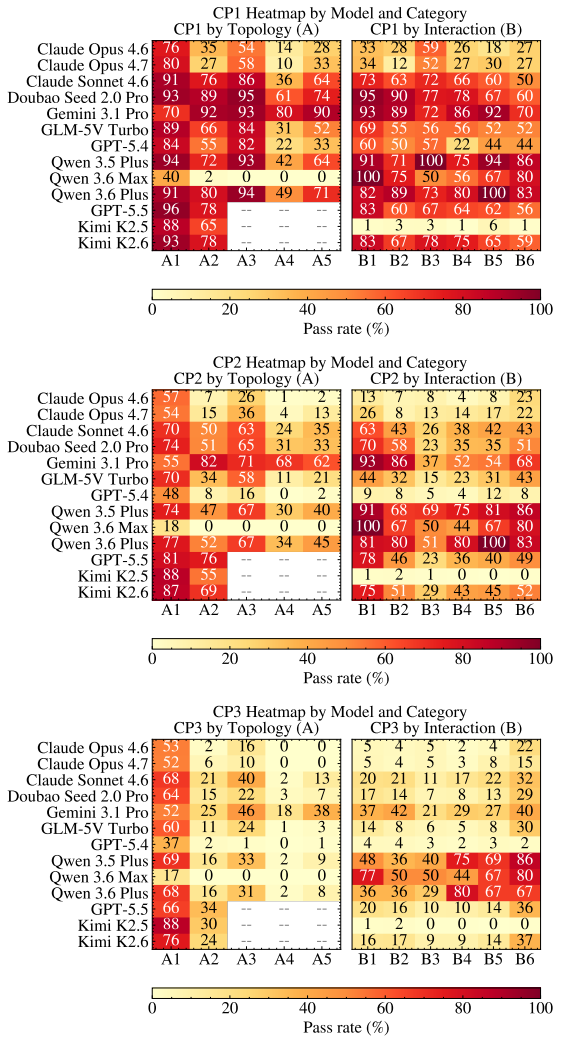


Figure 5: Heatmaps by model and category for CP1–CP3.

We will release our code, pipeline, and benchmark upon acceptance. Future work includes extending the benchmark to higher-loop orders and effective field theories, as well as evaluating the inverse generative task of synthesizing correct Feynman diagrams from textual physics descriptions.

## References


Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. 2024. UniBench: Visual Reasoning Requires Rethinking Vision-Language Beyond Scaling. arXiv:2408.04810 https://arxiv.org/abs/2408.04810

Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. 2024. MM-PhyQA: Multimodal Physics Question-Answering With Multi-Image CoT Prompting. arXiv:2404.08704 [cs.CL] doi:10.48550/arXiv.2404.08704

Anthropic. 2025a. Claude Opus 4.5 System Card. Anthropic System Card. https://www.anthropic.com/claude-opus-4-5-system-card Accessed: 2026-02-08.

Anthropic. 2025b. Claude Sonnet 4.5 System Card. Anthropic System Card. https://www.anthropic.com/claude-sonnet-4-5-system-card Accessed: 2026-02-08.

Ian Banta, Tianji Cai, Nathaniel Craig, and Zhengkang Zhang. 2024. Structures of Neural Network Effective Theories. *Phys. Rev. D* 109 (2024), 105007. arXiv:2305.02334 [hep-th] doi:10.1103/PhysRevD.109.105007

Jacob Biamonte. 2019. Lectures on quantum tensor networks.

Jacob D. Biamonte, Stephen R. Clark, and Dieter Jaksch. 2010. Categorical Tensor Network States. arXiv:1012.0531 https://arxiv.org/abs/1012.0531

Francesco Calisto, Ryan Moodie, and Simone Zoia. 2024. Learning Feynman integrals from differential equations with neural networks. *JHEP* 07 (2024), 124. arXiv:2312.02067 [hep-ph] doi:10.1007/JHEP07(2024)124

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A Theorem-driven Question Answering dataset. arXiv:2305.12524 [cs.CL] doi:10.48550/arXiv.2305.12524 Accepted to EMNLP 2023 (per arXiv record).

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. 2024. LLMPhy: Complex Physical Reasoning Using Large Language Models and World Models. arXiv:2411.08027 [cs.LG] doi:10.48550/arXiv.2411.08027

Alibaba Cloud. 2026. Visual Understanding (Qwen-VL) - Model Studio Documentation. https://www.alibabacloud.com/help/en/model-studio/vision.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Zi'ang Song, Guang Shi, and Haoqi Fan. 2025. Emerging Properties in Unified Multimodal Pretraining. arXiv:2505.14683 [cs.CV] doi:10.48550/arXiv.2505.14683 Introduces the open-source unified model BAGEL.

Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, et al. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences* 119, 32 (2022), e2123433119.

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. 2025. Unified Autoregressive Visual Generation and Understanding with Continuous Tokens. arXiv:2503.13436 [cs.CV] doi:10.48550/arXiv.2503.13436

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. 2025. PHYSICS: Benchmarking Foundation Models on University-Level Physics Problem Solving. arXiv:2503.21821 [cs.AI] doi:10.48550/arXiv.2503.21821

Richard P. Feynman. 1949. Space-Time Approach to Quantum Electrodynamics. *Physical Review* 76, 6 (1949), 769–789. doi:10.1103/PhysRev.76.769

Google Cloud. 2025. Gemini 3 Flash | Generative AI on Vertex AI. Google Cloud Documentation. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-flash Accessed: 2026-02-08.

Max Guillen, Philipp Misof, and Jan E. Gerken. 2025. Finite-Width Neural Tangent Kernels from Feynman Diagrams. arXiv. arXiv:2508.11522 [cs.LG] doi:10.48550/arXiv.2508.11522

Thomas Hahn. 2001. Generating Feynman Diagrams and Amplitudes with FeynArts 3. *Computer Physics Communications* 140, 3 (2001), 418–431. arXiv:hep-ph/0012260 doi:10.1016/S0010-4655(01)00290-9

Koji Hashimoto, Yuji Hirono, Jun Maeda, and Jojiro Totsuka-Yoshinaka. 2024. Neural network representation of quantum systems. *Machine Learning: Science and Technology* 5, 4 (2024), 045039. arXiv:2403.11420 [hep-th] doi:10.1088/2632-2153/ad81ac

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 3828–3850. doi:10.18653/v1/2024.acl-long.211

Yuji Hirono, Akinori Tanaka, and Kenji Fukushima. 2024. Understanding Diffusion Models by Feynman's Path Integral. arXiv. arXiv:2403.11262 [cs.LG] doi:10.48550/arXiv.2403.11262

Krešimir Kumerički. 2016. Feynman Diagrams for Beginners. arXiv:1602.04182 https://arxiv.org/abs/1602.04182

Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131* (2022).

David Leoni and Federico Franchini. 2024. Global sampling of Feynman's diagrams through normalizing flow. *Phys. Rev. Research* 6 (2024), 033041. arXiv:2402.00736 [hep-th] doi:10.1103/PhysRevResearch.6.033041

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. arXiv:2408.03326 https://arxiv.org/abs/2408.03326

Jindong Li, Yali Fu, Jiahong Liu, Linxiao Cao, Wei Ji, Menglin Yang, Irwin King, and Ming-Hsuan Yang.

2025. Discrete Tokenization for Multimodal LLMs: A Comprehensive Survey. arXiv:2507.22920 doi:10.48550/arXiv.2507.22920

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2310.03744 doi:10.1109/CVPR52733.2024.02484

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 https://arxiv.org/abs/2304.08485

Yuanche Liu, Yingxuan Xu, and Yang Zhang. 2025. Uncovering Singularities in Feynman Integrals via Machine Learning. arXiv. arXiv:2510.10099 [hep-ph] doi:10.48550/arXiv.2510.10099

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A Survey of Deep Learning for Mathematical Reasoning. 14605–14631 pages.

Harrison Mitchell, Alexander Norcliffe, and Pietro Liò. 2022. Learning Feynman Diagrams using Graph Neural Networks. arXiv. arXiv:2211.15348 [physics.comp-ph] doi:10.48550/arXiv.2211.15348 NeurIPS Machine Learning and the Physical Sciences (ML4PS), 2022.

Patrick Olivier. 2001. Diagrammatic reasoning: An artificial intelligence perspective. *Artificial Intelligence Review* 15, 1-2 (2001), 63–78. doi:10.1023/A:1006669526043

OpenAI. 2026a. GPT-5.1 Model | OpenAI API. OpenAI API Documentation. https://platform.openai.com/docs/models/gpt-5.1 Accessed: 2026-02-08.

OpenAI. 2026b. GPT-5.2 Model | OpenAI API. OpenAI API Documentation. https://platform.openai.com/docs/models/gpt-5.2 Accessed: 2026-02-08.

Antonio Pich. 2012. The Standard Model of Electroweak Interactions. arXiv:1201.0537 https://arxiv.org/abs/1201.0537

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. 2025. TokenFlow: Unified Image Tokenizer for Multimodal Understanding and Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2412.03069 [cs.CV] doi:10.48550/arXiv.2412.03069

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Arif Ahmed Sekh, Debi Prosad Dogra, Samarjit Kar, Partha Pratim Roy, and Dilip K Prasad. 2020. Can we automate diagrammatic reasoning? *Pattern Recognition* 106 (2020), 107412.

Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, Zhongwei Wan, Kai Zhang, Wendong Xu, Jing Xiong, Ping Luo, Wenhu Chen, Chaofan Tao, Zhuoqing Mao, and Ngai Wong. 2025. PhyX: Does Your Model Have the "Wits" for Physical Reasoning? arXiv:2505.15929 [cs.AI] doi:10.48550/arXiv.2505.15929

Vladyslav Shtabovenko, Rolf Mertig, and Frederik Orellana. 2016. New developments in FeynCalc 9.0. *Computer Physics Communications* 207 (2016), 432–444. arXiv:1601.01167 doi:10.1016/j.cpc.2016.06.008

Paolo Silvi, Florian Tschirsich, Mathias Gerster, Jasper Jünemann, Daniel Kasper, Miroslav Macek, and Simone Montangero. 2019. The Tensor Networks Anthology: Simulation Techniques for Quantum Many-Body Systems. *SciPost Physics Lecture Notes* 8 (2019). doi:10.21468/SciPostPhysLectNotes.8

Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. 2024. Multimodal Latent Language Modeling with Next-Token Diffusion. arXiv:2412.08635 [cs.CV] doi:10.48550/arXiv.2412.08635

David Tong. 2007. Lectures on Quantum Field Theory. https://www.damtp.cam.ac.uk/user/tong/qft.html Lecture notes, University of Cambridge (Michaelmas 2006).

Matt von Hippel and Matthias Wilhelm. 2025. Refining Integration-by-Parts Reduction of Feynman Integrals with Machine Learning. *JHEP* 05 (2025), 185. arXiv:2502.05121 [hep-th] doi:10.1007/JHEP05(2025)185

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, Mingyu Ding, Lei Bai, Wanli Ouyang, Shixiang Tang, Aoran Wang, and Xinzhu Ma. 2025b. PhysUniBench: An Undergraduate-Level Physics Reasoning Benchmark for Multimodal Models. arXiv:2506.17667 [cs.AI] doi:10.48550/arXiv.2506.17667

Weixing Wang, Zifeng Ding, Jindong Gu, Rui Cao, Christoph Meinel, Gerard de Melo, and Haojin Yang. 2025a. Image Tokens Matter: Mitigating

Hallucination in Discrete Tokenizer-based Large Vision-Language Models via Latent Editing. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:2505.21547 [cs.CV] doi:10.48550/arXiv.2505.21547

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Zhongyuan Wang, and Tiejun Huang. 2026. Multimodal learning with next-token prediction for large multimodal models. *Nature* (jan 2026). doi:10.1038/s41586-025-10041-x Online ahead of print (published 28 Jan 2026).

Xin Wang, Yuwei Zhou, Bin Huang, Hong Chen, and Wenwu Zhu. 2025c. Multi-modal Generative AI: Multi-modal LLMs, Diffusions and the Unification.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. 2025b. Harmonizing Visual Representations for Unified Multimodal Understanding and Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. arXiv:2503.21979 [cs.CV] doi:10.48550/arXiv.2503.21979

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. 2025a. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=02haSpO453

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. 2025. VisuLogic: A Benchmark for Evaluating Visual Reasoning in Multimodal Large Language Models. arXiv:2504.15279 https://arxiv.org/abs/2504.15279

Yan Yang, Haochen Tian, Yang Shi, Wulin Xie, Yi-Fan Zhang, Yuhao Dong, Yibo Hu, Liang Wang, Ran He, Caifeng Shan, et al. 2025. A Survey of Unified Multimodal Understanding and Generation: Advances and Challenges. TechRxiv preprint. doi:10.36227/techrxiv.176289261.16802577/v1 Posted on 11 Nov 2025.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. arXiv:2404.16006 https://arxiv.org/abs/2404.16006

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2311.16502 doi:10.1109/CVPR52733.2024.00913

Wanpeng Zhang, Yicheng Feng, Hao Luo, Yijiang Li, Zihao Yue, Sipeng Zheng, and Zongqing Lu. 2025. Unified Multimodal Understanding via Byte-Pair Visual Encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. arXiv:2506.23639 [cs.CV] doi:10.48550/arXiv.2506.23639

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. 2025. Unified multimodal understanding and generation models: Advances, challenges, and opportunities.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B. Tenenbaum, and Chuang Gan. 2024. ContPhy: Continuum Physical Concept Learning and Reasoning from Videos. arXiv:2402.06119 [cs.CV] doi:10.48550/arXiv.2402.06119

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI open* 1 (2020), 57–81.

Erle Zhu, Yadi Liu, Zhe Zhang, Xujun Li, Jin Zhou, Xinjie Yu, Minlie Huang, and Hongning Wang. 2025. MAPS: Advancing Multi-Modal Reasoning in Expert-Level Physical Science. arXiv:2501.10768 [cs.AI] doi:10.48550/arXiv.2501.10768