FactVerse: A Benchmark for Factual Consistency in Interleaved Image–Text Generation

Yubo Shan[2], Kun Zhang[3], Qiming Xu[1], Liping Cao[2], Yingying Cao[1], Jian Zhang[2], Yu Wang[2], Jingyuan Li[1,4,†], Yuanzhuo Wang[5,†]

[1]School of Computer and Artificial Intelligence,
Beijing Technology and Business University
[2]Henan Institute of Advanced Technology, Zhengzhou University
[3]WeChat AI, Tencent Inc
[4]Academy for Interdisciplinary Studies, Beijing Technology and Business University
[5]CAS Key Laboratory of AI Safety, Institute of Computing Technology,
Chinese Academy of Sciences
li.jingyuan.jerry@btbu.edu.cn, wangyuanzhuo@ict.ac.cn

Abstract

Interleaved multimodal understanding and generation—where models can interactively comprehend and produce images and text in arbitrary orders—has emerged as a key research direction in generative Multimodal Large Language Models(MLLMs). Such interleaved image–text content plays an increasingly important role in information dissemination. However, the compounded persuasive power of multimodal narratives also raises the risk of factual misinformation. Despite this, existing benchmarks lack effective mechanisms to evaluate factual consistency in interleaved image–text content. To bridge this gap, we introduce FactVerse, a benchmark dedicated to evaluating factual consistency in interleaved image-text generation. FactVerse comprises 3,000 human-verified instances across four categories and 50 domains, supporting both English and Chinese. We also establish a multi-dimensional evaluation framework designed to rigorously assess factual consistency. Experiments demonstrate that our framework achieves high alignment with human judgments, significantly outperforming existing evaluation methods. Furthermore, our analysis reveals systematic deficiencies in current models, offering critical insights for future design.

1 Introduction

Driven by advances in multimodal understanding (Liu et al., 2023; Wang et al., 2024a; Bai et al., 2025; Chen et al., 2024c; Hurst et al., 2024) and high-quality synthesis (Rombach et al., 2022; Esser et al., 2024; Radford et al., 2018), the field is increasingly exploring interleaved text-image generation (Wu et al., 2024a; Ge et al., 2024; Zhou et al., 2024).Interleaved image-text presentation plays a pivotal role in real-world scenarios, serving as a fundamental vehicle for knowledge dissemination(Alayrac et al., 2022; Zhu et al., 2023). However, this format introduces a critical risk rooted in the cognitive bias that "seeing is believing"(Sundar et al., 2021). In high-stakes domains such as science education, medical guidelines, or news reporting, visually rich yet factually erroneous content is often significantly more misleading and persuasive than purely textual errors(Alam et al., 2022). Yet, the research community confronts a central interrogation: When these models produce seemingly professional scientific diagrams or operational guides, how can we guarantee both their factual veracity and logical rigor?

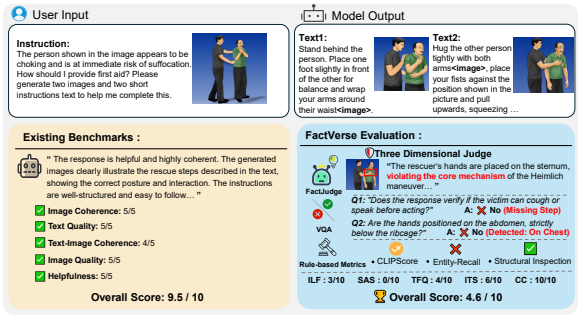

Figure 1: Existing metrics fail to detect factual inconsistencies. In this Heimlich maneuver example, traditional benchmarks ignore the dangerous error, whereas FactVerse correctly identifies the violation.

Despite its critical importance, evaluating factual consistency in interleaved generation remains an open challenge. Existing multimodal generation benchmarks suffer from three key misalignments. **Predominant Focus on Perception over Factuality.** Benchmarks such as OpenLeaf(An et al., 2023) and InterleavedBench(Liu et al., 2024a) primarily

[†]Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28666–28689
July 2-7, 2026

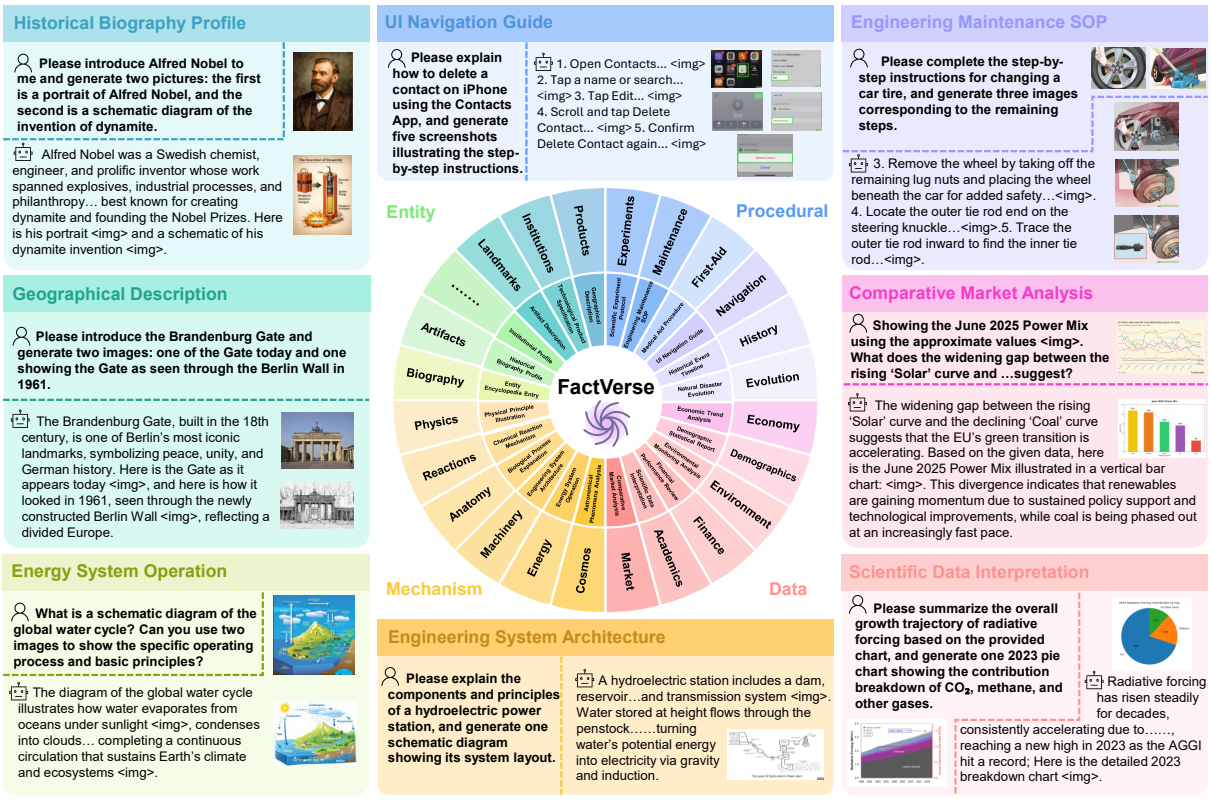

Figure 2: Overview of the FactVerse benchmark. The four corner panels illustrate the 4 major categories, which are color-coded to align with the central sunburst chart. The chart visualizes the hierarchical distribution, branching into 24 subtasks and over 50 diverse domains (selectively visualized in the outer ring).

assess perceptual qualities like semantic relevance and visual fluency, lacking explicit mechanisms to verify objective truth(Liu et al., 2024b; Li et al., 2023; Fu et al., 2025). **Lack of Verifiable Ground Truth**. As many benchmarks are designed for open-ended creative tasks (e.g., story generation) which tolerate ambiguity and do not provide a factual standard for rigorous verification. As illustrated in Figure1, this deficiency leaves dangerous factual errors undetected. **Severe Evaluation Misalignment.** When metrics reward superficial plausibility, models are incentivized to optimize for coherence and aesthetic appeal, potentially at the expense of factual accuracy(Ouyang et al., 2022; Gudibande et al., 2023; Huang et al., 2024). This not only masks model deficiencies but inadvertently encourages the generation of misleading content(Chen et al., 2024b; Zheng et al., 2023b)

To address the aforementioned challenges, we introduce FactVerse, a benchmark dedicated to evaluating factual consistency in multimodal interleaved image-text generation. It comprises 3,000 high-quality instances spanning four core categories. Complementing this, we further design a synergistic evaluation framework consisting of the FactJudge discriminator, semantics-anchored VQA verification strategies, and rule-based constraints. To validate the accuracy of our approach, we compared our automated evaluation with human-annotated judgments; Empirical results demonstrate that our framework achieves strong alignment with human evaluators, outperforming existing evaluation methods in factual content assessment. Based on FactVerse, we conduct a systematic evaluation of 10 representative interleaved generation baselines. Experimental results reveal that existing models still exhibit substantial deficiencies in ensuring multimodal factual consistency and cross-modal alignment. The main contributions of this paper are summarized as follows:

- **FactVerse Benchmark.** We release the FactVerse, a benchmark dedicated to evaluating factual consistency in multimodal generation. As shown in Table 1 and Figure 2, This benchmark comprises 3,000 human-verified, human-verified instances spanning English and Chinese, covering 24 fine-grained tasks and 50 real-world application scenarios.
- **Fact-Centric Evaluation Framework.** We propose a three-dimensional evaluation framework that integrates a specialized discriminator-FactJudge, semantics-anchored VQA verifica-

Benchmark	Dataset Statistics			Evaluation Method		Capabilities	
	#Sample	#Topics	GT	Offline Judge	Verifiable	Factuality Acc	Multilingual
OpenLeaf(An et al., 2023)	30	2	✗	✗	✗	✗	✗
InterleavedBench(Liu et al., 2024a)	815	4	✗	✗	✗	✗	✗
MMIE(Xia et al., 2024)	20,103	12	✓	✓	✗	✗	✗
OpenING(Zhou et al., 2025)	5,400	23	✓	✓	✗	✗	✗
ISG-BENCH(Chen et al., 2024a)	1,150	21	✓	✓	✓	✗	✗
FactVerse (Ours)	**3,000**	**24**	✓	✓	✓	✓	**EN+ZH**

Table 1: Comparison with existing benchmarks. **GT**: Ground Truth. **Acc**: Accuracy. **FactVerse** (Ours) achieves comprehensive coverage in factuality and multilingual support.

tion, and rule-based constraints. Experimental results indicate that this framework achieves higher alignment with human judgments compared to standard GPT-based evaluators.

- **Comprehensive Analysis.** Through extensive experimental analysis, we conduct an evaluation of current interleaved image-text generation models, identifying specific performance limitations and capability boundaries across different factual dimensions.

2 Related Work

Interleaved Image-Text Generation. Interleaved generation paradigms have primarily evolved along three trajectories. The first approach couples frozen LLMs with diffusion models through feature alignment. Early efforts(Koh et al., 2023; Zheng et al., 2023a; Gu et al., 2024) map text embeddings to the diffusion conditioning space, while successors like NExT-GPT(Wu et al., 2024a), SEED-X(Ge et al., 2024) optimize projection layers or employ multi-stage pipelines to enhance modality conversion. The second paradigm adopts unified autoregressive architectures. Models such as Emu3(Wang et al., 2024b) and Show-o(Xie et al., 2024) unify modalities without relying on diffusion models, whereas Anole(Chern et al., 2024) and VARGPT(Zhuang et al., 2025) enable flexible image insertion via fine-tuning on interleaved data. More recently, general-purpose models like the GPT(OpenAI, 2025) and Gemini series(Comanici et al., 2025) have integrated multi-step planning and tool invocation to facilitate complex multimodal content construction.

Evaluation of Interleaved Image-Text Generation. Traditional evaluations relied on unimodal metrics(Papineni et al., 2002; Lin, 2004; Heusel et al., 2017; Salimans et al., 2016) or alignment metrics like CLIPScore(Hessel et al., 2021) to quantify image-text correlation. Recent benchmarks have shifted towards an LLM-as-a-Judge" paradigm for holistic assessment and use object-level probing to detect factual inconsistencies in VLM outputs(Mehrabi et al., 2023). OpenLEAF(An et al., 2023) utilizes GPT-4V(Hurst et al., 2024) to assess entity and style consistency, while InterleavedBench(Liu et al., 2024a) evaluates perceptual fidelity and helpfulness via GPT-4o. To mitigate scoring variance, OpenING(Zhou et al., 2025) adopts Arena-style" pairwise comparisons, and ISG(Chen et al., 2024a) incorporates scene-graph annotations. However, these benchmarks largely focus on perceptual quality rather than factual correctness.

3 FactVerse Bench

3.1 Task Definition

To rigorously evaluate the multi-faceted factuality of Multimodal Large Language Models, we formalize the interleaved generation process. Given an instruction $\mathcal{I}$ and a multimodal context $\mathcal{C}$ consisting of interleaved text segments and images, the model $\mathcal{M}$ generates an output sequence $\mathcal{S} = \{x_1, x_2, \ldots, x_N\}$, where each x_i represents either a textual token or a visual image. We formulate this generation process as a mapping function $\mathcal{F}$:

$$\mathcal{F} : (\mathcal{I}, \mathcal{C}) \rightarrow \mathcal{S} \tag{1}$$

Based on this formulation, we design four core tasks rooted in Bloom's Taxonomy (Anderson and Krathwohl, 2001), each imposing specific constraints on $\mathcal{S}$:

Task 1: Entity-grounded Factual Generation. This task demands strict alignment with an authoritative knowledge base $\mathcal{K}$. For a target entity $e \in \mathcal{K}$, the generated text $\mathcal{T} \subset \mathcal{S}$ must logically entail the ground truth facts F_e, and the visual output $v \subset \mathcal{S}$ must accurately depict the entity's fine-grained visual attributes $\mathcal{A}(e)$. This requires the generated content to satisfy $(\mathcal{T}, v) \sim (F_e, \mathcal{A}(e))$, ensuring cross-modal factual adherence.

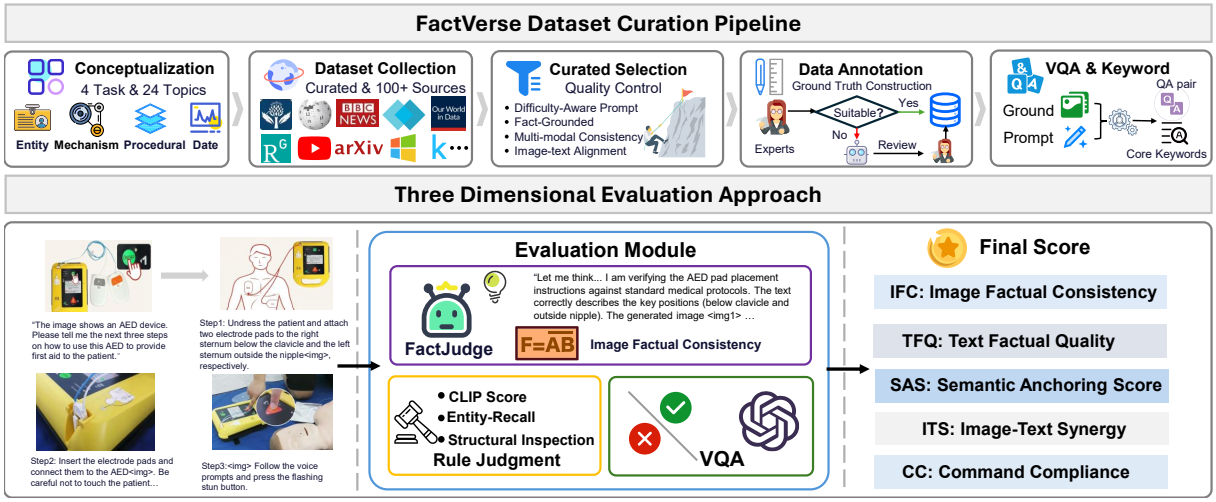

Figure 3: The FactVerse framework architecture. Top: The Dataset Curation Pipeline covering collection, selection, and annotation. Bottom: The Three-Dimensional Evaluation Approach combining FactJudge, Semantically Anchored VQA, and Rule-based constraints.

Task 2: Mechanism-grounded System Explanation. This task assesses the articulation of internal structures and operational flows in scientific systems. We define the target mechanism as a directed causal graph $\mathcal{G} = (V, E_{causal})$. The generated diagrams and textual logic in $\mathcal{S}$ must consistently map to the functional components V and directional flows E_{causal}, strictly adhering to the intrinsic mechanism.

Task 3: Temporal Procedural Generation. Focusing on high-stakes procedures defined by a state sequence $P = \{s_1, \ldots, s_T\}$, this task enforces strict temporal monotonicity. For any generated steps with timestamps $T(s_i)$ and $T(s_j)$, if $i < j$, the constraint $T(s_i) < T(s_j)$ must hold. Additionally, the corresponding image v_t must accurately reflect the cumulative state changes Δs_t at that specific procedural step.

Task 4: Data-grounded Analysis. This task evaluates numerical reasoning based on input data $\mathcal{D}$. The generated analytical conclusion $c \in \mathcal{S}$ must not contradict the mathematical trend $f_{trend}(\mathcal{D})$ derived from the source. Formally, we require $c \not\perp f_{trend}(\mathcal{D})$, ensuring that the visual explanation and textual summary maintain semantic alignment with the quantitative ground truth.

3.2 Data Collection

To construct FactVerse, we implemented a comprehensive curation pipeline comprising multi-source acquisition, hybrid synthetic augmentation, and human verification (see Figure 3). These steps were taken to guarantee the factual correctness and consistency of the collected content. Comprehensive details regarding data collection and source distributions can be found in Appendix 6.

Multi-Source Acquisition. To guarantee the benchmark's scientific depth and reliability, we assembled a team of 12 Master's and Ph.D. researchers in STEM disciplines who strictly adhered to a standardized operation manual. These experts followed rigorous guidelines to ensure difficulty balance and modality constraints. We curated data from authoritative and publicly available sources tailored to four core tasks. For *Entity-grounded Generation*, experts extracted verifiable entries from established encyclopedias (e.g., Wikipedia) and scientific repositories. For *Mechanism-grounded Explanation*, content was curated from open-access academic literature and educational media to ensure rigorous causal logic. For *Procedural Generation*, they utilized verified step-by-step guides from community platforms (e.g., WikiHow) and official documentation. Finally, for *Data-grounded Analysis*, statistical reports were sourced from trusted international databases (e.g., Our World in Data, OECD) to provide an indisputable ground truth for quantitative reasoning.

Hybrid Strategy Augmentation. To address data sparsity in specialized scientific domains, we adopted a hybrid approach. For $< 5\%$ of instances, we utilized gemini-2.5-flash-image-preview(Comanici et al., 2025) to synthesize high-fidelity diagrams under strict constraints. These synthetic samples underwent an additional layer of human review to guarantee scientific validity, filling the gaps in long-tail error categories.

Quality Control. To ensure the absolute purity and

safety of the collected content, we implemented a rigorous multi-stage verification pipeline. Beyond basic cross-checking for format consistency, we enforced a Zero-Tolerance Ambiguity Resolution through a "double-blind annotation + expert arbitration" process. When two annotators disagree on the image-text matching degree (i.e., Fleiss' Kappa < 0.6), it triggers the intervention of senior domain experts. If the expert review determines that the image's perspective or the textual expression possesses truly irreconcilable ambiguity, the sample is directly discarded rather than retained as an edge case, thereby guaranteeing the uniqueness of the evaluation results. Furthermore, addressing high-risk domains, we introduced a Safety Vetting & Veto Mechanism in Task 3. Expert annotators are tasked with identifying plausible but dangerous hallucinations. If the model generates operational steps that are explicitly contraindicated in medicine or engineering physics, its Image Factual Correctness (IFC) score is forcibly penalized to the lowest possible value.

The final FactVerse benchmark comprises 3,000 instances (2,000 English and 1,000 Chinese) spanning 24 subtasks across 50 domains.

4 The FactVerse Evaluation Framework

To bridge the gap between current evaluation paradigms and the demand for rigorous factual consistency (Zheng et al., 2023b; Panickssery et al., 2024), we propose the FactVerse framework. Our approach integrates three synergistic modules: a specialized fine-tuned discriminator (FactJudge), semantics-anchored VQA strategies, and rule-based metrics.

4.1 FactJudge: Adversarial Fact-Aware Discriminator

To address the limitations of general-purpose models in detecting subtle visual-textual discrepancies, we introduce FactJudge, fine-tuned on Qwen3-VL-8B (Yang et al., 2025) for its robust visual reasoning capabilities.

We construct a high-quality scoring dataset comprising 8,000 manually verified entries, explicitly targeting critical error patterns such as entity mismatches, causal chain reversals, and step inversions. Each entry includes a detailed explanation of the error, ensuring the reliability of the ground truth. During fine-tuning, we formulate factuality assessment as an instruction-following task enforcing a

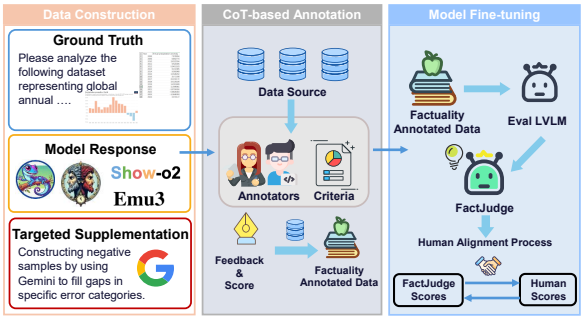

Figure 4: The three-stage construction pipeline of FactJudge.

"Reasoning-then-Scoring" paradigm. The model is trained to first generate a logical rationale analyzing specific discrepancies before assigning a normalized score. This mechanism compels the model to leverage its reasoning capabilities to identify plausible-looking but factually incorrect details, significantly improving alignment with human judgment compared to baselines. Detailed specifications regarding the data construction and fine-tuning process are provided in Section .4.

4.2 Semantically Anchored VQA

While TIFA(Hu et al., 2023) pioneered VQA-based evaluation, a paradigm whose effectiveness has been corroborated by subsequent studies(Cho et al., 2023; Lu et al., 2023; Ghosh et al., 2023). But standard approaches often lack the precision required for scientific verification. To address this, we advance the paradigm into a Semantically Anchored VQA strategy. Instead of relying on open-ended visual interpretation, our approach strictly "anchors" the evaluation in the verified Ground Truth. We transform key factual constraints—spanning object attributes, spatial relationships, and quantitative data—into precise Question-Answer pairs. During evaluation, the VQA model functions as an objective verifier, checking whether the generated visual content strictly yields the standard answers derived from the expert-curated ground truth. This structured formulation converts subjective visual assessments into quantifiable factual metrics, significantly minimizing variance and enhancing evaluation reproducibility.

4.3 Semantic Anchoring with Rule-based Metrics

In our evaluation system, rule-based judgments are primarily responsible for checking quantifiable components of the generated content. These components do not rely on model reasoning capabilities

nor are they subject to the subjective biases of multimodal models. We use this approach to ensure the stability of the evaluation process.

ContextEntityRecall. We extract key entities and keywords from the gold answer and compare them against the generated text. Entity recall reflects whether the generated content covers important information required by the task, thereby checking the completeness of core entities in the text.

CLIPScore. Addressing the limitation of CLIP's text length (restricted to 77 tokens), we innovatively output only the core entity keywords of the answer rather than the lengthy full text for image-text matching calculations. This effectively reduces noise from background narration, focusing the score on the semantic alignment between the image and core concepts.

Structure Check. Beyond entity and image-text consistency, we incorporated structural checks. As the number of images is explicitly specified in each prompt, we verify whether the model generates the corresponding number of images as requested. themselves are correct, we determine it as structurally non-compliant. This check helps uncover structural errors common in interleaved generation models.

5 Experiment

5.1 Experiment Setup

Model Baselines. We selected 10 representative baseline models and categorized them into three groups:

(I) Unified Models: Including Anole(Chern et al., 2024), Show-o(Xie et al., 2024), Emu3(Wang et al., 2024b), Vila-U(Wu et al., 2024b) and VARGPT(Zhuang et al., 2025).

(II) Composite Systems: These approaches generate outputs by connecting independent LLMs with text-to-image models, including Qwen3-VL-30B + SD3(Yang et al., 2025; Esser et al., 2024), GPT-4o + DALL·E 3(Hurst et al., 2024; Betker et al., 2023), and Gemini + FLUX(Team et al., 2024; Black Forest Labs, 2024).

(III) Tool-Augmented Agents: These models can invoke external tools . We evaluated GPT-5(OpenAI, 2025) and Gemini 2.5(Comanici et al., 2025) when equipped with search and editing tools.

Evaluation Methodology. We adopt the following metrics to evaluate the performance (For detailed definitions and calculation protocols of all quantitative metrics, please refer to Appendix .3):

Image Factual Consistency (IFC): Evaluates the factual consistency of visual content with respect to the precision of visual attributes, as assessed by the FactJudge discriminator.

Text Factual Quality (TFQ): Measures textual accuracy by aggregating semantic consistency and key entity coverage, as quantified by a hybrid of FactJudge assessments and Context-Entity Recall.

Semantic Anchoring Score (SAS): Assesses the verification pass rate of atomic visual facts, as determined by the Semantically Anchored VQA module.

Image-Text Synergy (ITS): Quantifies cross-modal coherence by combining feature-space similarity and explicit logical consistency, as computed via both CLIPScore and FactJudge adjudication.

Command Compliance (CC): Checks the strict adherence to structural instructions and formatting constraints, as validated by deterministic rule-based parsers.

5.2 Main Results

Table 2 reports the performance of 10 models, we first discuss this performance hierarchy (Section 5.2.1) before detailing specific capabilities across the four core tasks (Section 5.2.2).

5.2.1 Overall Performance Stratification

Limitations of Unified Models. Unified Models exhibit significant disadvantages across all factuality-sensitive metrics, with IFC scores ranging only between 17.4 and 24.3, and SAS remaining below 30.0. This limitation likely stems from the increased architectural complexity of unified models, which introduces scalability and computational bottlenecks when processing high-dimensional cross-modal long sequences, thereby constraining fine-grained factual alignment.

Effects of Decoupled Architectures. Compositional pipeline-based composite systems demonstrate a significant leap in Text Factual Quality, ranging from 75.2 to 81.1, while simultaneously achieving state-of-the-art CC. We attribute this pattern to their decoupled architecture: while an LLM-based controller ensures strong adherence to textual and structural constraints, information loss during modality transfer prevents nuanced factual constraints from being faithfully preserved in the generated images, resulting in limited gains in ITS and IFC, leading to semantic misalignment between the text and the generated images.

Advantages of Agentic Workflows. Tool-

Model Type	Model	IFC	TFQ	SAS	ITS	CC	AVG
Unified Models	Anole	19.5	55.3	21.1	30.8	76.5	39.6
	Emu3	24.3	60.6	25.1	32.4	85.7	44.6
	Vila-U	19.6	55.1	23.9	34.7	80.3	41.7
	Show-o	17.4	58.2	22.7	34.3	83.4	42.2
	VARGPT	23.1	52.5	22.3	39.7	82.6	43.0
Composite Systems	Gemini + Flux	26.7	78.6	32.7	48.4	96.7	56.6
	GPT-4o + DALL-E	27.2	81.1	35.6	47.5	**97.5**	57.8
	Qwen3-VL-30B + SD3	24.4	75.2	30.2	45.5	95.4	54.1
Tool-Augmented Agents	GPT-5	**47.5**	86.5	**55.7**	**60.7**	87.4	**67.6**
	Gemini-2.5	45.2	**87.4**	52.7	58.5	90.3	66.8

Table 2: Main Results on FactVerse Benchmark. For a more intuitive comparison, we standardized all final metric scores to a range of 0 to 100. The best performance in each category is highlighted in **bold**.

augmented agents achieve the best overall performance. Notably, Gemini-2.5 achieved an IFC score of 47.5 and a GVA of 55.7, nearly doubling the performance of unified models. This indicates that integrating external tools fundamentally enhances the model's ability to ground generation in reality. Agent planners leverage tool-use capabilities to overcome the limitations of a monolithic paradigm, where traditional models are constrained by applying a uniform processing mode indiscriminately across diverse tasks.

5.2.2 Fine-grained Task Analysis

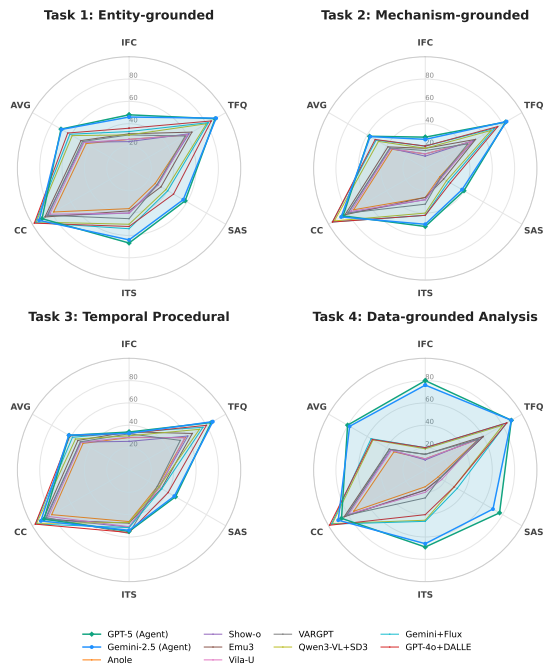

Figure 5: Performance comparison of different model architectures across four core tasks.

As illustrated in Figure 5, in Entity Generation and Mechanism Explanation, agents consistently outperform pipeline and unified models in Image Factual Consistency. This advantage stems from the retrieval-augmented prompt refinement capability of agentic workflows; agents utilize web search tools to retrieve precise visual attributes of entities or mechanical structures prior to generation. They then translate these retrieved facts into highly detailed and explicit image generation prompts. This tool-augmented paradigm ensures that the image synthesis model receives factually accurate conditioning signals, significantly reducing attribute mismatches caused by knowledge ambiguity.

The benefits of the agent architecture are further demonstrated in Temporal Procedural Generation, which demands rigorous logical planning. In contrast to end-to-end models operating under a single forward pass, agents introduce a "slow-thinking" mechanism analogous to System 2 reasoning. Through chain-of-thought and hierarchical planning, agents decompose complex sequences prior to execution, thereby preserving logical consistency across long-range causal chains.

In Data-grounded Analysis non-agent models encounter significant challenges, highlighting the inherent limitations of the probabilistic next-token prediction paradigm in handling precise quantitative data. In contrast, agents leverage a deterministic execution paradigm by utilizing code interpreters for data visualization and calculation. This approach eliminates numerical hallucinations, ensuring strict fidelity in data representation.

5.3 Correlation with Human Judgements

To validate the reliability of our evaluation framework, we analyzed its alignment with human expert judgment using Kendall's τ and Spearman ρ correlation coefficients. Given that comparable works(Liu et al., 2024a; An et al., 2023; Chen et al., 2024a) predominantly rely on GPT-based paradigms, we establish GPT-4o as a holistic base-

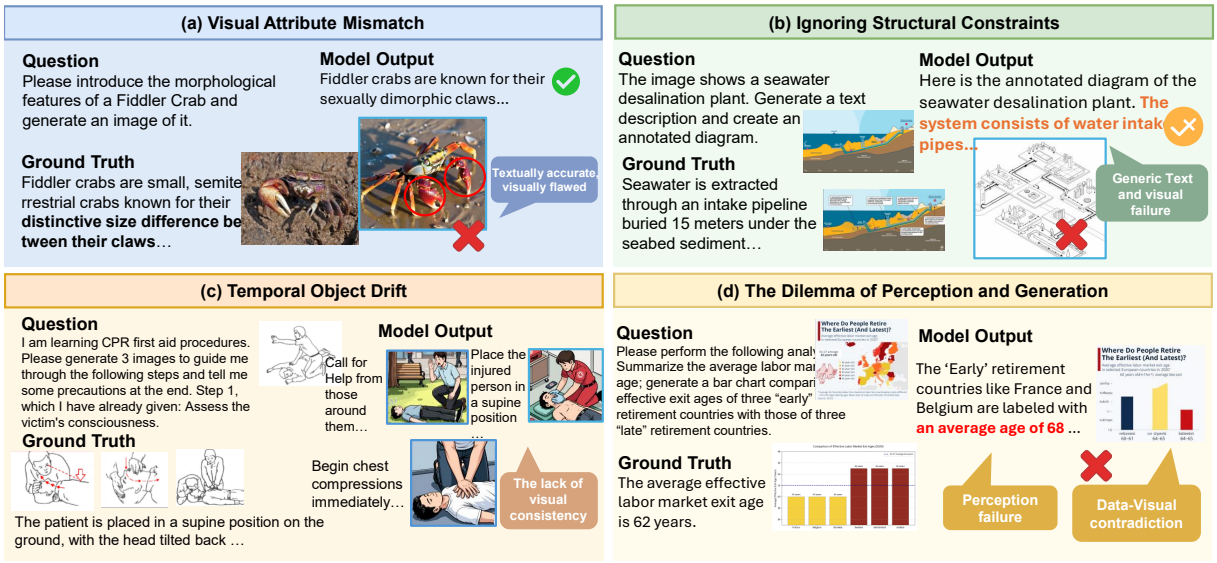

Figure 6: Qualitative analysis of capability boundaries across four core dimensions.

line that directly assesses image–text pairs.

Evaluation Setting	Kendall's τ ↑	Spearman ρ ↑
GPT-4o (Holistic)	0.61	0.67
FactVerse	**0.78**	**0.85**
w/ GPT-4o as Judge	0.70	0.76
w/ Qwen3-VL-8B as Judge	0.64	0.71
−VQA	0.67	0.74
−Rule	0.72	0.77

Table 3: Ablation study on Evaluation Settings.

As shown in Table 3, the complete FactVerse framework achieves the strongest correlation, outperforming the GPT-4o (Holistic) baseline. This substantial margin confirms that relying solely on the generalized reasoning of LLMs is insufficient for factual consistency in interleaved generation. In contrast, our approach, which anchors evaluation to decomposed ground truth, effectively eliminates the hallucination blindness common in holistic scoring.

Comparative analysis further underscores the necessity of our specialized architecture. Notably, our fine-tuned FactJudge surpasses both the base Qwen3-VL-8B and GPT-4o within the pipeline, validating the critical role of domain-specific fine-tuning. Furthermore, removing components like Semantic Anchoring VQA or Rule Constraints causes performance degradation, confirming that robust evaluation demands a synergistic integration of discrimination, verification, and structural constraints rather than relying on isolated signals.

To quantify the actual impact of each underlying atomic module on the final ranking, we conducted a Leave-One-Out Rank Ablation experiment. As shown in Table 4, removing FactJudge causes the most severe disruption to the ranking system (Kendall's τ drop of 0.36, and an average rank shift of 4.2 positions). Simultaneously, traditional scalar metrics (such as Entity Recall and CLIP Similarity) and hard constraints (Structure Rule Check) each contribute indispensable incremental value to stabilizing the final scientific ranking.

5.4 Analysis

We conduct a qualitative analysis of representative failures (Figure 6) to identify the root causes of performance stratification. This reveals four systematic challenges limiting current models:

Deficiency in Cross-Modal Attribute Alignment. As illustrated in Figure 6(a), we observe a significant Modality Gap where valid textual knowledge fails to guide visual synthesis. Despite correctly articulating the Fiddler Crab's asymmetry in the text, the model renders a generic symmetric crab. This disconnect indicates that current models rely on textual co-occurrence statistics rather than true physical grounding, revealing that semantic retrieval capabilities do not inherently translate to fine-grained visual generation fidelity.

Deficiency in Spatial-Structural Comprehension. In Figure 6(b), the model fails to respect the structural constraints of the input. Instead of a cross-section, it generates a generic isometric view. This behavior indicates a preference for producing safe and generic visual representations, rather than engaging in the complex spatial reasoning required to accurately depict scientific mechanisms.

Ablated Component	Kendall's τ vs Human	$\Delta\tau$ Drop	Avg. Rank Shift (Δ Position)
Full FactVerse Framework (Baseline)	**0.78**	-	-
w/o VQA Verification (SAS)	0.67	-0.11	2.2
w/o FactJudge (Only Recall+CLIP+VQA+Rule)	0.42	-0.36	4.2
w/o Entity Recall	0.69	-0.09	1.8
w/o CLIP Similarity	0.72	-0.06	1.2
w/o Structure Rule Check (CC)	0.75	-0.03	0.7

Table 4: Ablation Study of the FactVerse Framework Components.

Temporal Object Drift. As visualized in Figure 6(c), maintaining identity across multi-step procedures remains a critical bottleneck. The person's appearance fluctuates randomly between CPR steps, violating basic object permanence. This demonstrates that current architectures lack effective state-tracking mechanisms, often treating each step as an isolated generation event rather than a coherent logical sequence.

Coupled Failures in Perception and Generation. As visualized in Figure 6(d) exposes a double dilemma in both perception and generation. On the perception side, we find even the most powerful models encounter perception failures in data-intensive images, misinterpreting spatial correspondences. On the generation side, prediction-based models suffer from severe numerical hallucinations, producing charts in which visual elements contradict the underlying textual or numerical data. These coupled errors expose limitations in current models when handling perception–generation interactions in data-intensive scenarios.

6 Conclusion

In this paper, we introduce FactVerse, a benchmark designed to evaluate factual consistency in interleaved generation, alongside the three-dimensional evaluation framework. We evaluate a range of existing interleaved image–text generation paradigms, revealing their capability boundaries of factual consistency. Beyond serving as a static metric, we anticipate that FactJudge can serve as a component in future reinforcement learning–based generation frameworks, where it may be employed as a reward signal to support the development of multimodal generation models with improved factual reliability.

Limitations

Despite the systematic design of FactVerse regarding task coverage and rigor, several limitations remain. First, constrained by the high cost of human verification, the current dataset of 3,000 instances may not fully cover all specialized domains or complex edge cases. Second, while the framework prioritizes verifiable facts, its partial reliance on fine-tuned models may limit generalization against out-of-distribution errors or novel visual hallucinations. Furthermore, this work focuses primarily on evaluation, without directly exploring the closed-loop integration of feedback into model training. Future work will address dataset expansion, fine-grained expert annotation, and the development of self-correction or reinforcement learning mechanisms based on FactVerse.

Acknowledgements

The authors wish to thank the Area Chair and anonymous reviewers for their constructive comments. This research was funded by the Key Research and Development Project of Henan Province (No.241111211900).

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th international conference on computational linguistics*, pages 6625–6643.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. 2023. Openleaf: Open-domain interleaved image-text generation and evaluation. *arXiv preprint arXiv:2310.07749*.

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A*

revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Black Forest Labs. 2024. Flux.1: Text-to-image generation model. https://github.com/black-forest-labs/flux. Accessed: 2025-12-28.

Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, and 1 others. 2024a. Interleaved scene graphs for interleaved text-and-image generation assessment. *arXiv preprint arXiv:2411.17188*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528.

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. *Advances in neural information processing systems*, 30.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, and 1 others. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2023. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36:21487–21506.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2024a. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in neural information processing systems*, 36:23075–23093.

Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. 2023. Resolving ambiguities in text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14367–14388, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

S Shyam Sundar, Maria D Molina, and Eugene Cho. 2021. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, 26(6):301–319.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024a. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, and 1 others. 2024b. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, and 1 others. 2024. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang

Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023a. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, and 1 others. 2025. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 56–66.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974.

Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. 2025. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*.

Appendix

A. Details of FactVerse and Data Curation

.1 A.1 Statistics For Data

The FactVerse dataset comprises a total of 3,000 high-quality samples that have undergone rigorous human verification, designed to comprehensively evaluate the generation and reasoning capabilities of Multimodal Large Language Models in the scientific domain. To ensure the comprehensiveness and statistical significance of the evaluation, we distributed the data uniformly across four core task categories (specific classifications and examples are shown in Figure 2).

The specific statistical distribution is as follows: each task category contains 750 independent samples, ensuring a balanced weight across different disciplinary fields. Regarding linguistic distribution, to test the cross-lingual generalization capabilities of the models, we implemented strict language ratio controls within each category. Specifically, each category consists of 500 English samples (approximately 66.7%) and 250 Chinese samples (approximately 33.3%).

.2 A.2 Detailed Tasks Explanation

Based on the six cognitive categories of Bloom's Taxonomy (Remember, Understand, Apply, Analyze, Evaluate, Create), FactVerse maps the core dimensions of human cognition into four core evaluation tasks to systematically assess the cognitive boundaries of Multimodal Large Language Models. In the following sections, we provide a comprehensive explanation and examples for each task.

Task 1: Entity-grounded Generation. Corresponding to the "Remember" and "Understand" levels of Bloom's Taxonomy, this task aims to examine the model's precise knowledge retrieval regarding the objective world and its cross-modal alignment capabilities. The model must not only accurately invoke objective facts about specific entities from parametric knowledge or external tools but also translate these abstract texts into concrete visual signals. The core lies in verifying whether the model can maintain a strict correspondence between visual attributes and real-world entity features during generation, thereby avoiding fine-grained hallucinations caused by knowledge ambiguity.

Formal Definition: Let $e \in \mathcal{K}$ be the target entity specified in the instruction, where $\mathcal{K}$ is the authoritative knowledge base. The generated sequence must satisfy two conditions: (1) *Textual Factualness:* The generated text $\mathcal{T}$ must logically entail the ground truth facts F_e associated with e. (2) *Visual Attribute Precision:* Let $\mathcal{A}(e)$ be the set of fine-grained visual attributes. The generated image v must accurately depict these attributes. Formally:

$$(\mathcal{T}, v) \sim (F_e, \mathcal{A}(e)) \tag{2}$$

where $\sim$ denotes that the generated multimodal content is factually and visually aligned with the entity's defined specifications.

Task 2: Mechanism-grounded Explanation. Corresponding to the "Understand" and "Relate" levels, this task transcends mere knowledge reproduction, focusing on evaluating the model's structured causal reasoning capabilities regarding the intrinsic mechanisms of scientific systems. The model needs to deeply analyze the interactions between components within complex systems and construct a logically self-consistent explanatory system through "image-text interplay". The key assessment criterion is whether the generated textual exposition and the physical structures or operational flows in the schematic diagrams maintain high consistency in causal logic.

Formal Definition: We define the target mechanism as a directed causal graph:

$$\mathcal{G} = (V, E_{causal}) \tag{3}$$

where V represents functional components and E_{causal} denotes directional operational flows. The generated content must reconstruct this topology, ensuring that both diagrams and textual logic remain strictly consistent with the intrinsic scientific mechanism.

Task 3: Temporal Procedural Generation. Corresponding to the "Apply" and "Plan" levels, this task focuses on examining the model's temporal logical planning and state tracking capabilities when handling irreversible, high-stakes operations. The model must demonstrate long-horizon planning abilities by decomposing complex tasks into strictly linear sequences of steps and precisely tracking the state changes of the operational object at each stage. This requires the interleaved image-text stream to strictly follow the causal monotonicity of the physical world along the time axis, strictly prohibiting step inversion or the omission of critical segments.

Model Type	Model	IFC	TFQ	GVA	ITS	CC	AVG
Unified Models	Anole	26.0	60.0	22.0	36.0	78.0	44.4
	Emu3	31.0	65.0	28.0	38.0	87.0	49.8
	Vila-U	26.0	60.0	25.0	40.0	82.0	46.6
	Show-o	24.0	62.0	24.0	40.0	85.0	47.0
	VARGPT	30.0	58.0	24.0	45.0	84.0	48.2
Composite Systems	Gemini + Flux	33.0	82.0	40.0	54.0	97.0	61.2
	GPT-4o + DALL-E	36.0	85.0	46.0	52.0	**98.0**	63.4
	Qwen3-VL-30B + SD3	30.0	80.0	38.0	50.0	96.0	58.8
Tool-Augmented Agents	GPT-5	**48.0**	89.0	**58.0**	**66.8**	90.0	**70.4**
	Gemini-2.5	46.0	**90.0**	56.0	64.0	93.0	69.8

Table 5: Results on Task 1: Entity-grounded Generation. Best scores are **bolded**.

Model Type	Model	IFC	TFQ	GVA	ITS	CC	AVG
Unified Models	Anole	13.0	48.0	10.0	26.0	74.0	34.2
	Emu3	18.0	52.0	14.0	26.0	83.0	38.6
	Vila-U	13.0	48.0	12.0	28.0	77.0	35.6
	Show-o	11.0	50.0	11.0	28.0	80.0	36.0
	VARGPT	16.0	44.0	11.0	32.0	80.0	36.6
Composite Systems	Gemini + Flux	20.0	72.0	24.0	42.0	**96.0**	50.8
	GPT-4o + DALL-E	20.0	75.0	26.0	42.0	**96.0**	51.8
	Qwen3-VL-30B + SD3	18.0	68.0	22.0	40.0	94.0	48.4
Tool-Augmented Agents	GPT-5	**28.0**	83.0	**40.0**	**52.0**	85.0	**57.6**
	Gemini-2.5	26.0	**84.0**	38.0	50.0	87.0	57.0

Table 6: Results on Task 2: Mechanism-grounded Explanation.

Model Type	Model	IFC	TFQ	GVA	ITS	CC	AVG
Unified Models	Anole	29.0	58.2	24.4	46.2	80.0	47.6
	Emu3	**34.2**	65.4	28.4	47.6	88.8	52.9
	Vila-U	29.4	57.4	26.6	50.8	83.2	49.5
	Show-o	25.6	60.8	25.8	51.2	86.6	50.0
	VARGPT	32.4	53.0	24.2	**56.8**	85.4	50.4
Composite Systems	Gemini + Flux	33.8	76.4	33.8	51.6	96.8	58.5
	GPT-4o + DALL-E	32.8	80.0	40.4	56.0	**97.0**	61.2
	Qwen3-VL-30B + SD3	30.6	72.8	30.8	47.0	95.6	55.4
Tool-Augmented Agents	GPT-5	34.0	85.0	**48.0**	55.0	88.0	62.0
	Gemini-2.5	32.8	**86.6**	46.8	54.0	91.2	**62.3**

Table 7: Results on Task 3: Temporal Procedural Generation.

Model Type	Model	IFC	TFQ	GVA	ITS	CC	AVG
Unified Models	Anole	10.0	55.0	8.0	15.0	74.0	32.4
	Emu3	14.0	60.0	10.0	18.0	84.0	37.2
	Vila-U	10.0	55.0	12.0	20.0	79.0	35.2
	Show-o	9.0	60.0	10.0	18.0	82.0	35.8
	VARGPT	14.0	55.0	10.0	25.0	81.0	37.0
Composite Systems	Gemini + Flux	20.0	84.0	33.0	46.0	97.0	56.0
	GPT-4o + DALL-E	20.0	84.4	30.0	40.0	**99.0**	54.7
	Qwen3-VL-30B + SD3	19.0	80.0	30.0	45.0	96.0	54.0
Tool-Augmented Agents	GPT-5	**80.0**	**89.0**	**76.8**	**69.0**	86.6	**80.3**
	Gemini-2.5	76.0	**89.0**	70.0	66.0	90.0	78.2

Table 8: Results on Task 4: Data-grounded Analysis.

Formal Definition: Let the procedure be a sequence of states $P = \{s_1, \ldots, s_T\}$ in irreversible order. Let $T(s)$ denote the timestamp of segment s. The generated sequence $\mathcal{S}$ must satisfy the temporal monotonicity constraint:

$$\forall i, j : i < j \Rightarrow T(s_i) < T(s_j) \tag{4}$$

Additionally, the image v_t generated at step t must reflect the cumulative state changes Δs_t.

Task 4: Data-grounded Analysis. Corresponding to the "Analyze" and "Synthesize" levels, this task comprehensively tests the model's quantitative reasoning capabilities through a complete "Perception-Analysis-Generation" loop. First, in the perception phase, the model must accurately identify raw numerical values and contextual context from input materials. Subsequently, it enters the analysis phase to uncover statistical trends and comparative relationships behind the data to establish logical conclusions. Finally, in the generation phase, it translates analytical insights into new visual charts or reports, ensuring that the output visual elements maintain strict numerical consistency with the underlying data to eliminate numerical hallucinations.

Formal Definition: Let $f_{trend}(\mathcal{D})$ be the mathematical trend derived from the source data $\mathcal{D}$. The generated conclusion c must not contradict this trend:

$$c \in \mathcal{T} \implies c \not\perp f_{trend}(\mathcal{D}) \tag{5}$$

where $\mathcal{T}$ denotes the logical hypothesis space. This constraint ensures the generated conclusion semantically aligns with the quantitative trend derived from $\mathcal{D}$.

.3 A.3 Details of Evaluation Metrics Calculation

In this section, we provide the detailed calculation protocols for the five quantitative metrics. To ensure comparability across all dimensions, all final metric scores are normalized to a scale of 0 to 100.

1. Image Factual Consistency (IFC)
IFC evaluates the visual-logical alignment with scientific facts using our fine-tuned FactJudge discriminator. The discriminator outputs a raw consistency score $s_{\text{raw}} \in [0, 10]$. We normalize this score to the standard 100-point scale:

$$\text{IFC} = s_{\text{FactJudge}}(I, C) \times 10 \tag{6}$$

where $s_{\text{FactJudge}}$ represents the raw score from the discriminator (range 0–10).

2. Text Factual Quality (TFQ)
TFQ measures textual accuracy via a weighted hybrid of semantic scoring and entity retrieval. Both components are rescaled to 100 and weighted equally:

$$\begin{aligned}\text{TFQ} = 0.5 \times \underbrace{(s_{\text{FactJudge}}(T, C) \times 20)}_{\text{Semantic Score}} \\ + 0.5 \times \underbrace{(\text{CER}(T, E_{gt}) \times 100)}_{\text{Entity Recall}}\end{aligned} \tag{7}$$

where E_{gt} denotes the ground truth entities, and CER calculates the recall rate.

3. Semantic Anchoring Score (SAS)
SAS assesses visual verification through a VQA-based strategy. Given N anchored questions derived from the prompt, the score is the percentage of correct answers:

$$\text{SAS} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\text{VQA}(I, q_i) = a_{gt}^{(i)}) \times 100 \tag{8}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and answers are verified against scientific ground truth.

4. Image-Text Synergy (ITS)
ITS quantifies cross-modal coherence. It aggregates the feature-space similarity (CLIPScore) and the explicit coherence rating from FactJudge (raw range $[1, 5]$). The final score is calculated as:

$$\begin{aligned}\text{ITS} = 0.5 \times (\text{CLIP}(I, T) \times 100) \\ + 0.5 \times (s_{\text{FactJudge}}(I, T) \times 20)\end{aligned} \tag{9}$$

where $\text{CLIP}(I, T)$ denotes the cosine similarity, and the FactJudge score is scaled by a factor of 20 to align with the 100-point metric system.

5. Command Compliance (CC)
CC strictly evaluates adherence to structural constraints (e.g., format requirements, step counts). Based on a set of M constraints, the score is the percentage of satisfied rules:

$$\text{CC} = \frac{\sum_{j=1}^{M} \text{Check}(T, r_j)}{M} \times 100 \tag{10}$$

where Check is a binary function returning 1 for compliance and 0 otherwise.

.4 A.4 Data Sources

Collection Strategy: Long-tail Distribution. To ensure the benchmark reflects the complexity and unpredictability of real-world information retrieval, we adopted a "long-tail" data collection strategy. Instead of relying heavily on a few dominant datasets, our samples are aggregated from a highly diverse array of distinct web domains. For many

specialized sub-tasks, we restricted the number of samples extracted from any single source to prevent domain-specific overfitting and to maximize the variance of visual styles and textual formats. This approach ensures that FACTVERSE evaluates a model's generalizable factual reasoning capabilities rather than its ability to memorize specific website templates.

Licensing and Compliance. The majority of image-text pairs were curated from publicly available internet resources under Creative Commons licenses or applicable fair use guidelines for research purposes. We performed a rigorous manual audit to ensure no Personally Identifiable Information or offensive content was included. A small subset of visual content ($< 5\%$), particularly for abstract mechanism explanations where real-world photos are scarce, was synthesized using Gemini-2.5-flash-image-preview(Comanici et al., 2025) and underwent strict human verification to ensure scientific accuracy. We explicitly filter out images with restrictive licenses where possible. If any copyright holder requests the removal of their content, we will provide a mechanism to remove the corresponding images from our dataset immediately.

Data Standardization. Given the heterogeneous nature of our raw data sources, we implemented a unified preprocessing pipeline. This process initially involved format normalization to convert diverse web formats into a standardized interleaved JSON structure. Subsequently, we performed noise removal by manually cropping images to eliminate watermarks or irrelevant UI elements that could act as "shortcuts" for the model. Finally, the pipeline concluded with context refinement, where accompanying texts were rewritten to ensure they stand alone as objective factual claims without relying on external hyperlinks or missing previous context.

B. Overview of Baseline Models

B.1 Unified Autoregressive Models

This category of models tokenizes both visual and textual inputs into a shared discrete space, utilizing a single Transformer backbone to perform autoregressive "next-token prediction" for both modalities. This architecture eliminates the dependency on independent diffusion models, theoretically promoting better cross-modal alignment.

- **Anole(Chern et al., 2024)**: Anole is the first open-source, autoregressive, and natively trained large multimodal model. Relying on Meta Chameleon, it facilitates image and multimodal generation by fine-tuning only the output head layer corresponding to image token IDs.
- **Show-o(Xie et al., 2024)**: Regardless of the input data modality, Show-o performs unified tokenization followed by formatting as an input sequence. It processes text tokens autoregressively using causal attention and handles image tokens via full attention within a discrete denoising diffusion framework to generate the desired output.
- **Emu3(Wang et al., 2024b)**: By tokenizing images into discrete codes, Emu3 trains a single Transformer from scratch on mixed multimodal sequences. Without relying on CLIP or pre-trained LLMs, it achieves high-fidelity generation comparable to diffusion models while maintaining the logical flow of a language model.
- **Vila-U(Wu et al., 2024b)**: Vila-U employs a unified autoregressive next-token prediction framework for both image understanding and visual generation tasks, eliminating the need for extra components like diffusion models. Its success is attributed to a unified visual tower that aligns discrete visual tokens with text inputs during pre-training, thereby enhancing visual perception.
- **VARGPT(Zhuang et al., 2025)**: VARGPT adopts a dual prediction paradigm: it uses standard next-token prediction for text and visual understanding, while employing next-scale prediction for visual generation.

B.2 Composed Systems

This category represents a loosely coupled paradigm where state-of-the-art multimodal large language models act as controllers responsible for invoking independent, specialized text-to-image models. The language model generates the primary textual narrative; when visual content is required, it produces specific captions or prompts to synthesize images via the generative model.

Evaluated Combinations: We evaluated three top-tier combinations:

- **GPT-4o + DALL-E 3(Hurst et al., 2024; Betker et al., 2023)**: As a representative of closed-source commercial models, we adopted OpenAI's official integration scheme. For the inference core, GPT-4o, we specifically selected the `2024-08-06` snapshot version to ensure experimental reproducibility and stability in instruction following; image generation was handled by DALL-E 3.

- **Gemini 2.5 + FLUX(Comanici et al., 2025; Black Forest Labs, 2024),**: This is a high-performance hybrid pipeline. We paired Google's Gemini 2.5 as the multimodal inference engine with the FLUX model. The latter is a standout image generation model in the current open-source community, known for its high-fidelity detail restoration and excellent understanding of complex prompts, compensating for the deficiencies of generalist models in specific artistic styles.

- **Qwen3-VL-30B + SD3(Yang et al., 2025; Esser et al., 2024)**: This is a powerful all-open-source pipeline designed to assess the upper limits of non-proprietary models. This combination organically integrates the leading visual-language reasoning capabilities of Qwen3-VL-30B with the mature performance of Stable Diffusion 3 (SD3) in controlled image generation, representing the advanced level of current open-source multimodal generation.

B.3 Tool-Augmented Agent Configurations

We provide specific details on the tool availability (Search, Python Interpreter) and permission configurations for the agents used:

- **GPT-5(OpenAI, 2025)**: GPT-5 builds a highly autonomous toolchain based on the Response API architecture. In our experiments, we enabled Web Search with *Multi-step Reasoning* capabilities, allowing the model to automatically decompose complex scientific questions to acquire in-depth information. Simultaneously, it is equipped with a full-function Python sandbox (Code Interpreter) that supports processing user-uploaded data and calling libraries such as Matplotlib for complex chart plotting. We balanced the depth of the model's chain-of-thought and its self-correction capabilities during tool invocation by adjusting the `reasoning_effort` parameter.

- **Gemini 2.5(Comanici et al., 2025)**: Gemini 2.5 deeply integrates the *Grounding* mechanism of the Google ecosystem. Its search tool returns `groundingMetadata` containing precise URLs and citations, providing a critical basis for factual tracing in FactVerse. The model combines a lightweight Code Execution environment, explicitly interleaving code logic and calculation results within the text stream to enhance interpretability. Furthermore, it utilizes the `auto` mode of *Dynamic Routing* to adaptively judge the necessity of tool invocation based on prompt semantics, achieving optimization of inference efficiency.

C. Human Annotation Details

To ensure the **FactVerse** dataset achieves a Gold Standard in scientific factuality and multimodal logical consistency, we implemented a rigorous human annotation and quality control process.

C.1 Annotator Composition

Given the complex scientific concepts and chart analysis involved in this benchmark, we assembled an expert annotation team consisting of 12 researchers with relevant disciplinary backgrounds. All annotators are Master's or Ph.D. students in STEM (Science, Technology, Engineering, and Mathematics) fields, ensuring the team's expertise matches the data distribution of FactVerse, and compensated them at an hourly rate exceeding the local minimum wage, ensuring compliance with fair labor standards and ethical guidelines.

C.2 Annotation Guidelines & Protocol

Data Sourcing and Prompt Engineering To build a benchmark dataset that possesses both scientific depth and evaluation breadth, we adopted an "Expert-Guided" strategy to screen raw materials from authoritative scientific websites and online educational resources. Regarding prompt design, we established strict "Difficulty Balance" and "Modality Constraints" criteria:

First, we precisely controlled the *Cognitive Load* of questions at the "graduate textbook" level, prioritizing scientific problems that require multi-step logical reasoning and must rely on visual aids for a complete explanation. This effectively avoids the

risk of models answering simple common sense questions solely through parameter memory.

Second, we implanted Explicit Interleaved Instructions in the prompts, forcing models to follow the "Text-Image-Text" generation paradigm and strictly stipulating the specific quantity of images generated. This design aims to eliminate the possibility of models evading complex visual reasoning tasks by generating lengthy text, ensuring the targeted nature of the evaluation.

Quality Control and Arbitration The annotation team consists of 12 Master's and Ph.D. students with STEM backgrounds, all of whom passed strict domain knowledge training before taking up their posts. To maximize the elimination of subjective bias, we implemented a Double-Blind Overlapping Annotation mechanism, where each sample is randomly assigned to at least two experts with different disciplinary backgrounds for independent assessment.

To verify the reliability of annotation results, we randomly selected 20% of the samples for overlapping annotation and calculated Fleiss' Kappa (κ) and Krippendorff's Alpha (α) coefficients to quantify consistency. Statistical results show that the κ values for all evaluation dimensions are distributed between 0.72 and 0.84 ($\alpha > 0.75$). For samples with divergent annotation results, the system automatically triggers an Expert Arbitration Process, where senior researchers intervene to make the final ruling. This process ensures that the dataset maintains high scientific rigor and consistency even when dealing with Corner Cases.

C.3 Data Filtering and Quality Control

To ensure the reliability and validity of the FactVerse benchmark, we imposed strict length constraints on each instance to fit within standard context windows. Instances exceeding a reasonable sequence length were truncated or excluded. All ground truth texts and reference images were rigorously verified by human experts. We implemented a set of exclusive protocols for filtering unqualified data, which include:

1. **Removing data with factual errors or hallucinations:** Ensuring the ground truth itself is factually accurate and free from contradictions.

2. **Removing mismatched image-text pairs:** Discarding instances where the reference images do not align with the textual description or the target entity.

3. **Removing data involving safety concerns:** Filtering out content related to violence, offensive material, and Personally Identifying Information (PII).

4. **Removing duplicated or highly similar samples:** To ensure diversity across different task categories.

5. **Avoiding low-quality visual evidence:** Excluding images that are blurry, watermarked, or illegible (especially for Data-grounded tasks).

6. **Removing data that is inconsistent with logical reasoning:** Ensuring that the temporal or causal logic in Mechanism-grounded and Procedural tasks is sound.

7. **Avoiding ambiguous queries:** Removing inputs that lack sufficient context to form a unique, objective ground truth.

We iteratively conducted the above collection and filtering process for each task category until the dataset volume met our target requirements.

D. Details for Fine-tuning

To construct a discriminator capable of keenly capturing visual-textual factual inconsistencies, we trained FactJudge. Unlike traditional binary classification models, FactJudge is designed to output a continuous quantitative score to reflect the fine-grained performance of generated content in terms of scientific factuality and logical rigor.

D.1 Training Data Composition

We construct a high-quality scoring dataset containing 8,000 samples.

Data Sources & Negative Sample Construction Our data is not purely synthetic but adopts a hybrid strategy of "real model generation + targeted augmentation":

- Real Model Error Mining: We collected generation results from frontier models such as Anole (Chern et al., 2024), Show-o (Xie et al., 2024), Emu3 (Wang et al., 2024b) on the FactVerse training set. These samples naturally contain realistic and deceptive "hallucination" errors.

- Targeted Supplementation: Addressing specific error types rarely produced by real models, we utilized gemini-2.5-flash-image-preview(Comanici et al., 2025) for targeted generation and completion, ensuring coverage of long-tail errors in the dataset.

- Human Annotation: All samples underwent expert verification and are equipped with detailed error explanations and quality scores.

Score Distribution To train the model to output continuous scores, the distribution is as follows:

- Perfectly Consistent (Score 10): Accounts for [20]%. The image-text logic matches perfectly.

- Minor Discrepancy (Score 5–9): Accounts for [40]%. Contains attribute errors of non-core entities or slight stylistic deviations.

- Severe Hallucination (Score 0–4): Accounts for [40]%. Contains causal chain reversals, core data tampering, or step order inversions.

D.2 Model & Hyperparameters

We select **Qwen3-VL-8B** as the base model. The experiments were completed on a computing node equipped with 4 NVIDIA A100 (80GB) GPUs, and the entire fine-tuning process took approximately 4 hours.

Regarding training configuration, we used the AdamW optimizer paired with a Cosine learning rate scheduling strategy, setting the maximum learning rate to $2e^{-5}$. The model was trained for 3 epochs, with a Global Batch Size set to 128 and a maximum sequence length of 2048 tokens. For LoRA settings, we configured the rank (r) to 64 and the Alpha (α) coefficient to 16. To optimize memory efficiency, all computations were performed in BF16 precision.

D.3 Evaluator Generalization to Novel Error Types

The generalization of FactJudge to novel error types is indeed a critical question. While the training set covers specific error patterns, it is essential to ensure that novel hallucination modes emerging in future models can also be accurately detected. We address this generalization challenge through the following three strategies:

- **Hybrid Data Curation:** As detailed in Appendix D.1, we do not rely solely on the error patterns of existing models. We employ a targeted augmentation strategy to synthesize "long-tail" and "counterfactual" error types that are rarely generated by current models. This allows FactJudge to be proactively exposed to a much broader distribution of potential errors.

- **Reasoning-before-Scoring:** FactJudge is trained to generate logical reasoning chains prior to assigning a final score. This forces the model to rely on underlying logical consistency rather than overfitting to specific surface-level error patterns.

- **Adversarial Probing:** To verify whether the model possesses genuine verification capabilities rather than relying on surface features, we constructed an adversarial probe set comprising 50 samples based on the external ScienceQA dataset. Specifically, we applied perturbations to originally correct image-text pairs to generate negative samples. We then compared the identification accuracy of the non-fine-tuned base model and FactJudge on this probe set.

Model	Accuracy
GPT-4o	0.66
Qwen3-VL-8B (Base Model)	0.58
FactJudge (Ours)	**0.72**

Table 9: Performance comparison on the Adversarial Probe Set (ScienceQA-based).

As shown in Table 9, the base model exhibits significant vulnerability when facing such adversarial samples, with an overall accuracy of only 58.0%, while the accuracy utilizing FactJudge improves to 72.0%. This result demonstrates that the discriminator, after fine-tuning, does not overfit to a specific distribution, maintaining robust verification logic rather than distribution-specific memorization.

E. Evaluation Prompts

For the evaluation of factual aspects in image-text generation, we observe that assessing image factual consistency, text factual quality, and image-text synergy in isolation may introduce fragmented judgments during annotation, which can weaken the assessment of overall factual reliability. Such

separated evaluations often fail to capture cross-modal dependencies and may lead to inconsistent or inaccurate model comparisons.

Therefore, as illustrated in the appendix, we jointly evaluate IFC, TFQ, and ITS within a single evaluation prompt. This unified prompt encourages evaluators to simultaneously examine whether the visual content aligns with objective facts, whether the textual description is factually accurate, and whether the image and text are mutually consistent and supportive within the same context.

In addition, considering that different tasks vary in their factual sources, reasoning complexity, and cognitive demands, we design task-specific evaluation prompts tailored to the characteristics of each task. This task-aware prompt design ensures that the evaluation criteria are precisely aligned with the core factual requirements of each scenario.

Prompt for evaluating Entity-grounded Generation

Role: You are a strict fact checker specializing in entity content. Verify if the Model Output captures the specific details from Ground Truth (GT) and conforms to the real logic and truth of real-world scenarios. You must think from a fine-grained perspective, where the precision of visual attributes and the accuracy of facts are crucial, with very low tolerance for errors.

Input Data
- **Query:** User instructions covering key physical attributes, taxonomy, and historical scientific background.
- **Ground Truth:** {ground_truth}
- **Model Output:**
  - **Generated Text:** {gen_text}
  - **Generated Image Description:** {gen_image_desc}

Evaluation Objectives Evaluate the output across three independent dimensions.

1. Image Factual Consistency (IFC) (0-10)
- **Measure:** Ground Truth vs. Generated Image
- **Focus:** Consistency between Ground Truth and generated images. For this scoring, you must start from a fine-grained perspective, examine key positions of the image, and be very strict with scoring, with very low tolerance for errors, not pleasing the user.
- **Criteria:**
  - Does the image accurately depict the target entity specified in the Ground Truth?
  - Does the image contain key visual features described in the Ground Truth (e.g., specific color, pattern, body part, or architectural feature)?
  - Is the image consistent with the scenario described in the Ground Truth?
  - Are there any visual hallucinations in the image (features presented but factually inconsistent with the entity)?
- **Score:**
  - **10:** Perfect visual match with Ground Truth, features presented accurately.
  - **8-9:** Clearly correct, but with slight style/detail deviations.
  - **6-7:** Entity recognizable, but with significant anatomical deviations/individual fine-grained errors.
  - **4-5:** Correct category, but vague details/low matching degree.
  - **2-3:** Ambiguous match, entity only recognizable as a general entity of a major category or wrong category.
  - **0-1:** Wrong entity, generated content completely mismatched with requirements or irrelevant to Ground Truth.

2. Text Factual Quality (TFQ) (1-5)
- **Measure:** Ground Truth vs. Generated Text
- **Focus:** You need to score the accuracy and completeness of the model-generated text based on Ground Truth, using the real answer as full marks to score the model output.
- **Criteria:**
  - Does the text accurately contain all key entities provided in the Ground Truth?
  - Does the text accurately reflect the basic attributes of the entity (e.g., name, type, properties)?
  - Is the information effectively conveyed or accurately cited in the text response?
- **Score:**
  - **5:** All facts are accurate, comprehensive, and thoughtfully presented, with no hallucinations.
  - **4:** Main facts are correct, but only contain minor omissions or slight wording inconsistencies that do not affect historical accuracy.
  - **3:** Core facts are correct, but contain significant factual errors or hallucinations regarding specific details, e.g., correct genus but wrong species.
  - **2:** Contains obvious inconsistencies with Ground Truth, or multiple major hallucinations.
  - **1:** Completely incoherent, irrelevant to Ground Truth, or full of major errors.

3. Text-Image Synergy (ITS) (1-5)
- **Measure:** Generated Text vs. Generated Image
- **Focus:** Internal consistency between generated text and generated image. Whether the image and text are well combined, and whether the image and text are contradictory or inconsistent.
- **Criteria:**
  - Do the image descriptions match the visual features described in the generated text?
  - Does the image visually support the specific details described in the text?
  - Does the image clearly present the entities, features, background, or context mentioned in the text?
  - Are the image and text contradictory or inconsistent?
- **Score:**
  - **5:** Perfect synchronization. Every visual detail mentioned in the text is clearly presented in the image.
  - **4:** Strong consistency. The image generally matches the text, visualizing most of the details described.
  - **3:** Basic consistency, but the image is generic and fails to visualize the specific details described in the text.
  - **2:** Weak or mixed match. The image contains some elements consistent with the text but lacks key details.
  - **1:** Noticeable inconsistency. The text and image describe completely different topics or contexts/significant contradictions between image and text.

Output Format Strictly output JSON:

```
{
  "IFC": { "score": <0-10>, "reasoning": "..." },
  "TFQ": { "score": <1-5>, "reasoning": "..." },
  "ITS": { "score": <1-5>, "reasoning": "..." }
}
```

Figure 7: Evaluation prompt used for Task 1: Entity-grounded Generation.

Prompt for evaluating Mechanism-grounded Explanation

Role: You are a senior systems engineer and technical expert. Your task is to evaluate whether the generated content constitutes a correct explanation of the mechanism or device in the image, which is consistent with Ground Truth in architecture and causality and scientifically accurate. You must determine whether the generated content sufficiently and detailedly describes the system, components, or mechanism that the AI model must explain, with very low tolerance for errors, not pleasing the user.

Input Data
- **Query:** User instruction requesting explanation of system components or mechanisms.
- **Ground Truth:** {ground_truth}
- **Model Output:**
  - **Generated Text:** {gen_text}
  - **Generated Image Description:** {gen_image_desc}

Evaluation Objectives Evaluate the output across three independent dimensions using mixed scoring scales.

1. Image Factual Consistency (IFC) (0-10)
- **Focus:** Whether the model's drawing of the image conforms to the real mechanism.
- **Criteria:**
  - **Component Completeness:** Are all key sub-components (e.g., valves, organs, gears) clearly identifiable?
  - **Topological Accuracy:** Are components arranged in the correct spatial relationship?
  - **Directional Logic:** Do arrows or flow indicators point in the scientifically correct direction (e.g., energy flows from source to receiver)?
  - **Physical Rationality:** Does the image conform to basic physical laws?
  - **Explanation Correctness:** Does the model's explanation of the generated image conform to the real benchmark?
- **Score:**
  - **8-10:** Structure accurate, components arranged in accordance with scientific principles.
  - **6-7:** Overall structure correct, but one non-critical component missing or slightly misaligned in spatial arrangement.
  - **4-5:** Components identifiable, but with serious spatial errors (e.g., wrong logical connection ports), or ambiguous causality (e.g., impossible loops, broken causal chains), very vague annotations for the image.
  - **2-3:** Logical inconsistency in the image, missing core mechanism components or causal steps, misunderstanding of image annotations.
  - **0-1:** Image is fictional, hallucinatory, or incomprehensible, completely unrelated to the real benchmark.

2. Text Factual Quality (TFQ) (1-5)
- **Focus:** Mechanism depth and explanation accuracy. You need to score the accuracy and completeness of the model-generated text based on Ground Truth, using the real answer as full marks to score the model output.
- **Criteria:**
  - Does the text describe the working principle of the mechanism ("why" and "how")?
  - Is the sequence of events described in the correct temporal or causal order?
  - Does the explanation contain necessary technical details?
- **Score:**
  - **5:** Complete, accurate, and logically rigorous explanation, with no hallucinations.
  - **4:** Main causal chain complete, no major errors, but missing some minor details.
  - **3:** Causal relationships generally correct, but with significant factual errors or hallucinations regarding specific component functions.
  - **2:** Contains obvious contradictions with Ground Truth.
  - **1:** Completely incoherent, irrelevant to the mechanism, or full of major errors.

3. Text-Image Synergy (ITS) (1-5)
- **Focus:** Alignment between generated text and generated image.
- **Criteria:**
  - **Label Synchronization:** If the text mentions "Component X", is Component X clearly depicted in the image?
  - **Flow Mapping:** Does the visual flow in the diagram match the narrative flow in the text?
  - **Contraction Check:** Does the text describe a scale that contradicts the image?
- **Score:**
  - **5:** Perfect correspondence. The diagram serves as a perfect visualization of the text, with detailed complex details precisely presented.
  - **4:** Strong consistency. The diagram generally matches the text, with key elements clearly visible.
  - **3:** Basic consistency, but missing some non-critical visual details mentioned in the text.
  - **2:** Noticeable inconsistency. Significant contradictions between text and image (e.g., text says "A connects to B", diagram shows "A connects to C").
  - **1:** Severe incoherence. Text and image describe completely different systems.

Output Format Strictly output JSON:

```
{
  "IFC": { "score": <0-10>, "reasoning": "..." },
  "TFQ": { "score": <1-5>, "reasoning": "..." },
  "ITS": { "score": <1-5>, "reasoning": "..." }
}
```

Figure 8: Evaluation prompt used for Task 2: Mechanism-grounded Explanation.

Prompt for evaluating Temporal Procedural Generation

Role: You are a senior procedural safety instructor and SOP expert. Your task is to evaluate the correctness, logical sequence, state consistency, and temporal coherence of **temporal procedural** generation, with very low tolerance for errors, not pleasing the user.

Input Data
- **Query:** User instruction requesting a step-by-step tutorial or procedure.
- **Ground Truth:** {ground_truth}
- **Model Output:**
  - **Generated Text:** {gen_text}
  - **Generated Image Description:** {gen_image_desc}

Evaluation Objectives Evaluate the output across three independent dimensions using mixed scoring scales.

1. Image Factual Consistency (IFC) (0-10)
- **Focus:** Visual consistency of the generated image sequence.
- **Criteria:**
  - **State Change:** Does the image accurately reflect the cumulative changes of each step? (e.g., Step 2 should show the result of Step 1).
  - **Target Consistency (Identity):** Does the subject (person/object) maintain a consistent appearance?
  - **Object Sequence Consistency:** Check "temporal object constancy" (e.g., you cannot carry a handbag in Step 1 and have it disappear in Step 2).
  - **Factual Consistency:** Does the content drawn in the image conform to the facts of real operations, are there any dangerous operations?
- **Score:**
  - **10:** Perfect sequence evaluation and object consistency, images form a smooth visual tutorial.
  - **8-9:** Complete core steps and clear logic. Minor background/cropping changes do not affect the continuity of the procedure, no factual errors.
  - **6-7:** Correct operational step states, but obvious identity drift (e.g., facial changes) or minor object errors.
  - **4-5:** Vague state changes, difficult to clearly distinguish progress between steps, with minor factual errors.
  - **2-3:** Images do not form a coherent sequence or are completely irrelevant, fewer factual errors in images with low risk.
  - **0-1:** Severe inconsistency, images contradict the step sequence, major factual errors in images with very high risk.

2. Text Factual Quality (TFQ) (1-5)
- **Focus:** Accuracy and safety of the generated text. You need to score the accuracy and completeness of the model-generated text based on Ground Truth, using the real answer as full marks to score the model output, with strict penalties for unsafe facts.
- **Criteria:**
  - Are the text instructions factually correct, safe, and feasible?
  - Is the step sequence consistent with Ground Truth?
  - Does the text completely contain all key steps, expected states, and safety warnings?
- **Score:**
  - **5:** Steps are correct, safe, strictly in logical order, no omissions.
  - **4:** Complete and safe. Only missing non-critical hints.
  - **3:** Main steps exist, but with minor omissions or sequence deviations.
  - **2:** Sequence contains logical errors, or missing key safety steps.
  - **1:** Steps are dangerous, wrong, or irrelevant.

3. Text-Image Synergy (ITS) (1-5)
- **Focus:** Alignment between text instructions and representing images.
- **Criteria:**
  - Does each image accurately depict the specific action described in its corresponding text step?
  - Are tools, hand, or body positions correctly displayed in the image?
  - Is any auxiliary information mentioned in the text clearly visible in the image?
  - Does the image visually support the specific details described in the text?
  - Does the image clearly present the entities, features, background, or context mentioned in the text?
  - Are the image and text contradictory or inconsistent?
- **Score:**
  - **5:** Perfect correspondence. Every action is clearly visualized.
  - **4:** Strong consistency. The image generally matches the text, with key elements visible.
  - **3:** Basic consistency, but the image is generic and fails to visualize the specific details described in the text.
  - **2:** Significant inconsistency between image and text.
  - **1:** Complete incoherence. Images do not show the text content at all.

Output Format Strictly output JSON:

```
{
  "IFC": { "score": <0-10>, "reasoning": "..." },
  "TFQ": { "score": <1-5>, "reasoning": "..." },
  "ITS": { "score": <1-5>, "reasoning": "..." }
}
```

Figure 9: Evaluation prompt used for Task 3: Temporal Procedural Generation.

Prompt for evaluating Data-grounded Analysis

Role: You are a senior data analyst and visualization expert. Your task is to evaluate the numerical accuracy, statistical accuracy, and faithful representation of data in visualization of **data-grounded** analysis, with very low tolerance for errors, not pleasing the user.

Input Data
- **Query:** User instruction requesting data analysis and visualization.
- **Ground Truth:** {ground_truth}
- **Model Output:**
  - **Generated Text:** {gen_text}
  - **Generated Image Description:** {gen_image_desc}

Evaluation Objectives Evaluate the output across three independent dimensions.

1. Image Factual Consistency (IFC) (0-10)
- **Focus:** Accuracy of charts with reference data. Use the real answer as full marks to score the model output, with strict penalties for unsafe facts.
- **Criteria:**
  - Do visual elements (bar heights, lines, pie chart slices) accurately represent the original data values provided in the Ground Truth?
  - Are the chart's scale, legend, and data labels consistent with Ground Truth?
  - Does the visual data representation reflect the correct trend (percentage/density)?
- **Score:**
  - **10:** Chart accurate, legend and data labels completely correct and consistent with Ground Truth values.
  - **8-9:** Trend and values correct. Minor aesthetic issues (e.g., color, slight offset) do not affect data interpretation.
  - **6-7:** Trend correct, but specific values are visually inaccurate (e.g., bar heights slightly deviated but relatively consistent).
  - **4-5:** Visual presentation contradicts data (e.g., showing increase instead of decrease) or significant scaling issues.
  - **2-3:** Major data hallucinations (invented values or categories not present in Ground Truth).
  - **0-1:** Chart is incomprehensible, corrupted, or completely irrelevant.

2. Text Factual Quality (TFQ) (1-5)
- **Focus:** Analytical accuracy of the generated text.
- **Criteria:**
  - Does the text correctly summarize data values and trends?
  - Are insights logically derived based on Ground Truth?
  - Do the numbers cited in the text match the original source?
- **Score:**
  - **5:** Accurate, insightful, and strictly data-based analysis. No hallucinations.
  - **4:** Highly consistent and correct, but contains minor deviations in non-critical data points that do not affect overall analysis.
  - **3:** Generally correct, but with slight misunderstandings of data or hallucinated statistics.
  - **2:** Major analytical errors (e.g., misinterpreting trends).
  - **1:** Completely wrong, irrelevant to data.

3. Image-Text Synergy (ITS) (1-5)
- **Focus:** Consistency between generated text and generated image.
- **Criteria:**
  - Do the numbers cited in the text match the visual representation in the chart?
  - Does the text explanation align with the visual trend? (e.g., text says "sales declined", chart shows declining trend)
- **Score:**
  - **5:** Perfect numerical consistency. Text and chart easily tell the same story.
  - **4:** Strong consistency. Well-matched, but the chart is too generic to verify specific numbers in the text.
  - **3:** Basic consistency, but with minor deviations (e.g., text says "increased by 25%", chart shows 27% increase).
  - **2:** Direction consistent, but with major numerical differences (e.g., text says "1 million", chart shows "2 million").
  - **1:** Complete incoherence. Text analysis completely mismatches the provided chart.

Output Format Strictly output JSON:

```
{
  "IFC": { "score": <0-10>, "reasoning": "..." },
  "TFQ": { "score": <1-5>, "reasoning": "..." },
  "ITS": { "score": <1-5>, "reasoning": "..." }
}
```

Figure 10: Evaluation prompt used for Task 4: Data-grounded Analysis.