GENIUS: GENERATIVE FLUID INTELLIGENCE EVALUATION SUITE

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified Multimodal Models (UMMs) have shown remarkable progress in visual generation. Yet, existing benchmarks predominantly assess *Crystallized Intelligence*, which relies on recalling accumulated knowledge and learned schemas. This focus overlooks *Generative Fluid Intelligence (GFI)*: the capacity to induce patterns, reason through constraints, and adapt to novel scenarios on the fly. To rigorously assess this capability, we introduce **GENIUS** (**GEN**erative Fluid **I**ntelligence Eval**U**ation **S**uite). We operationalize *GFI* through three complementary dimensions. These include *Inducing Implicit Patterns* (e.g., inferring personalized visual preferences), *Executing Ad-hoc Constraints* (e.g., visualizing abstract metaphors), and *Adapting to Contextual Knowledge* (e.g., simulating counter-intuitive physics). Collectively, these dimensions challenge models to solve problems whose governing rules are established by the immediate context. Our systematic evaluation of 14 representative models reveals significant performance deficits in these tasks. Our diagnostic analysis reveals a gap between recognizing the intended outcome and faithfully generating it: models can identify the expected output in a discriminative VQA probe, yet struggle to execute contextual rules during generation. To bridge this gap, we propose a training-free attention intervention strategy. Ultimately, **GENIUS** establishes a rigorous standard for *GFI*, guiding the field beyond knowledge utilization toward dynamic, general-purpose reasoning.

1 INTRODUCTION

Unified Multimodal Models (UMMs) have witnessed remarkable progress recently (Team, 2024; Chen et al., 2025; Xie et al., 2024), delivering impressive results across diverse tasks (An et al., 2025; Li et al., 2025a; Jiang* et al., 2025). Benefiting from the fusion of understanding, UMMs are capable of processing complex, interleaved contexts and exhibiting extensive world knowledge to reshape the generative paradigm. Consequently, they are widely seen as a milestone toward Artificial General Intelligence (AGI). However, this rapid advancement invites a natural question: *How far are current UMMs from true general intelligence regarding visual generation?*

To investigate this problem, drawing upon existing literature (Cattell, 1963; Schipolowski et al., 2014; Kent, 2017), we deconstruct General Intelligence in visual generation into two primary components: *Crystallized Intelligence (CI)* and *Fluid Intelligence (FI)*. Current development and evaluation focus of UMMs mainly targets *CI* (i.e., the capacity for memorization and retrieval of pre-trained knowledge). For instance, a model's ability to generate a flawless "cat" often stems from exposure to billions of instances during training, followed by probabilistic reproduction during inference. However, this trend has severely masked a critical but long-ignoring deficiency concerning *FI* of visual generation skills, termed *Generative Fluid Intelligence (GFI)*, (i.e., the ability to perform inducing, reasoning and ad-hoc adaptation in novel scenarios). As shown in Figure 1, the "Simple Constraint" task requires identifying ad-hoc rules (e.g., abstract symbol denotes "rain") and applying them, rather than retrieving static concepts.

Despite its critical importance, research along this direction remains limited (shown in Table 1):

- **First, an operational framework is needed.** A task-oriented formulation of GFI can connect induction, rule execution, and contextual adaptation to concrete requirements in visual generation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

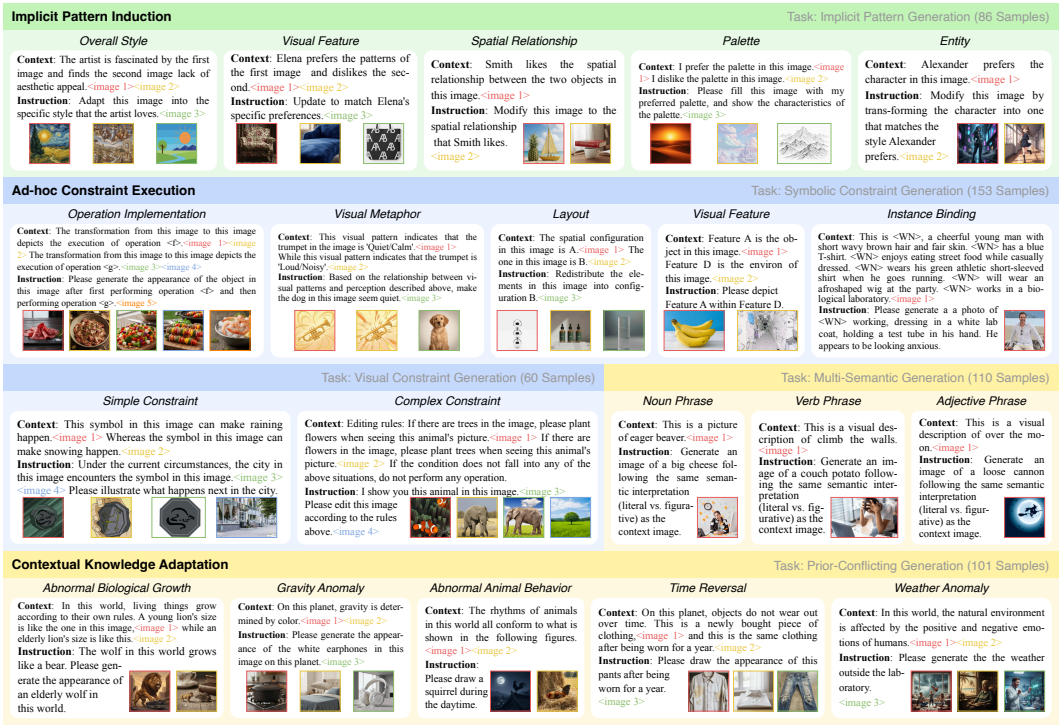


Figure 1: **GENIUS overview.** It's structured into three dimensions, five tasks and diverse sub-tasks.

- **Second, benchmark coverage remains limited.** Evaluating GFI requires tasks in which the immediate context determines the target rule, together with metrics that assess whether the generated output follows that rule.
- **Third, systematic analyses are lacking.** The lack of investigations into the failure modes leaves critical questions of *why models fail* and *how to improve* unanswered.

To bridge these gaps, we introduce **GENIUS** (**GEN**erative Fluid **I**ntelligence Eval**U**ation **S**uite), the first framework dedicated to the systematic evaluation of GFI. Drawing from the Cattell-Horn-Carroll (CHC) theory (Schneider & McGrew, 2012), we distill three core dimensions of *FI*: (I) *Inductive Inference*, (II) *Abstract Dynamic Reasoning* and (III) *Adaptive Inhibition*. We materialize these theoretical concepts into three corresponding dimensions within **GENIUS**. To ensure a fine-grained assessment, we further construct five novel and well-designed tasks that specify concrete capabilities within each dimension. For each task, we employ a hybrid evaluation comprising three metrics: (I) *Rule Compliance*, which challenges the model's precision in following ad-hoc rules; (II) *Visual Consistency*, which assesses the stability of generated attributes under logical constraints; and (III) *Aesthetic Quality*, which demands that the model maintains fundamental aesthetic standards. Through manual curation, our suite features tasks well designed by multi-modal experts. Unlike traditional benchmarks that prioritize static world knowledge, generation quality or safety, we construct tasks in which the intended transformation must be determined from the supplied examples, temporary definitions, or contextual knowledge, reducing opportunities to succeed through pretrained knowledge.

With **GENIUS**, we systematically evaluate 14 representative open-source, proprietary and newly proposed agentic generation models. To ensure evaluation robustness, we provide manually annotated hints for each test case, which have undergone at least three rounds of cross-validation to support unbiased hybrid evaluation. Overall, our results reveal clear gaps between current state-of-the-art (SOTA) models and general intelligence. Surprisingly, pre-planning and post-reflection yield marginal gains. These findings expose under-explored deficiencies in current generative models, highlighting the urgent need to advance fluid intelligence in the next generation of UMMs.

Building on these findings, we move beyond evaluation to investigate the underlying mechanisms of failure. Taking Bagel (Deng et al., 2025) as an example, by visualizing the attention distribution of it, it surprisingly shows irregular noise and spikes across the multimodal context. This indicates that the model struggles to accurately focus on critical new rules in the context. Inspired by the

Table 1: **Comparison of representative benchmarks.** * denotes understanding tasks. ✔ indicates partial satisfaction (e.g., For Manual Curated/Annotation, it implies a combination of human curation and automatic methods). GENIUS pioneers Fluid Intelligence evaluation, featuring multi-image input, multimodal interleaved context, hybrid metrics, and manually curated testcases.

Benchmark	# Samples	Multi-Image Input	Fluid Intelligence	Multimodal Interleaved Context	Task Dimension			Hybrid Evaluation	Manual Curated Annotation
					Implicit Pattern Induction	Explicit Constraint Execution	Contextual Knowledge Adaptation		
GenEval (Ghosh et al., 2023)	2.2k	✗	✗	✗	✗	✗	✗	✗	✗
WISE (Niu et al., 2025)	1.0k	✗	✗	✗	✗	✗	✗	✗	✔
RISE (Zhao et al., 2025)	360	✗	✗	✗	✗	✗	✗	✓	✓
DPG-Bench (Hu et al., 2024)	1.0k	✗	✗	✗	✗	✗	✗	✗	✗
OpenING* (Zhou et al., 2025)	5.4k	✗	✗	✗	✗	✗	✗	✗	✔
UniEval* (Li et al., 2025e)	4.2k	✗	✗	✗	✗	✗	✗	✗	✗
MME-Unify* (Xie et al., 2025b)	4.1k	✓	✗	✗	✗	✗	✗	✗	✔
RealUnify* (Shi et al., 2025)	1.0k	✓	✗	✗	✗	✗	✗	✗	✓
ROVER (Liang et al., 2025)	1.3k	✓	✗	✗	✗	✗	✗	✓	✓
WEAVE (Chow et al., 2025)	0.1k	✓	✗	✔	✗	✗	✗	✓	✓
GENIUS (Ours)	0.5k	✓	✓	✓	✓	✓	✓	✓	✓

theoretical perspective of *In-Context Learning (ICL) as Implicit Fine-Tuning* (Dherin et al., 2025), we demonstrate that ICL process is mathematically equivalent to update specific model parameters while generation. Then, we offer a possible view: the imbalanced attention distribution results in insufficient guidance during implicit gradient descent, which causes the gradient direction vague or stochastic, failing to overcome pre-trained priors. Based on this, we design a training-free mechanism as a strong baseline. The results show consistent performance gains across all tasks.

In summary, our core contributions are as follows:

- We operationalize *Generative Fluid Intelligence (GFI)*, filling a theoretical void to provide foundational guidance for steering UMMs toward general intelligence.
- We introduce **GENIUS**, the first benchmark designed to systematically and purely quantify *GFI*. It features 510 carefully expert-curated samples spanning three dimensions and five tasks, supported by a robust hybrid evaluation protocol with well designed and human-written rubrics.
- We systematically evaluate 14 representative open-source, proprietary and agentic generation models. Results reveal significant deficits in SOTA models, underscoring clear performance gap.
- We trace *GFI* failures via theoretical and empirical analysis. Then, we propose a training-free strategy that boosts performance, effectively activating the model's *GFI*.

2 GENIUS

2.1 Benchmark Overview

Grounded in the Cattell-Horn-Carroll (CHC) theory (Schneider & McGrew, 2012), General Intelligence is defined not merely by *Crystallized Intelligence (CI)* but more fundamentally by *Fluid Intelligence (FI)*, which is the capacity to induce, reason, and adapt in novel scenarios, independent or even contrary to prior knowledge. Building on this, we operationalize *FI* into an actionable and diagnostically framework specifically for visual generation. We introduce *Generative Fluid Intelligence (GFI)*, structured along three core dimensions that directly map to the established narrow abilities within CHC theory:

- **Implicit Pattern Induction (maps to Inductive Reasoning)**: The ability to discover implicit patterns. In visual generation, this requires extracting "soft" unstated rules—such as implicit user preferences or unnameable aesthetic styles—from raw, unstructured multimodal observations.
- **Ad-hoc Constraint Execution (maps to Dynamic Reasoning):** The capacity to execute newly defined rules. This involves applying logical reasoning within "hard" ad-hoc constraints (e.g., executing visual operations or abstract symbols defined solely in the immediate context).
- **Contextual Knowledge Adaptation (maps to Adaptive Inhibition):** The capability to suppress stale priors. When immediate contextual cues conflict with pretrained knowledge (e.g., physical laws), the model must actively inhibit its crystallized priors to faithfully adapt to the current context.

These dimensions distinguish how the governing rule is obtained and applied: inferred from examples, executed through temporary definitions, or followed when contextual evidence conflicts with pretrained expectations. This decomposition offers an intuitive and empirically diagnosable framework tailored for visual generation. It systematically distills the complex generative process into three

foundational steps: determining what latent structures to extract (Induction), formulating how to execute the newly defined rules (Execution), and resolving conflicts with intrinsic priors (Adaptation). To rigorously quantify these capabilities, **GENIUS** is designed to isolate external priors, ensuring evaluation is driven primarily by the immediate interleaved context rather than pre-memorized priors.

2.2 Benchmark Construction

For each category of **GENIUS**, we curate a diverse set of high-quality, expert-designed test cases. Each instance comprises a multi-modal interleaved context and removal of any single modality from context makes the task unsolvable.

Implicit Pattern Induction: This dimension mainly contains a novel task: *Implicit Pattern Generation*, which assesses the capacity to deduce unstated visual preferences from context and apply them for generation. As shown in Figure 1, the interleaved input presents images of varying styles alongside specific user preferences. During testing, the model is required to induce the target stylistic pattern based on these preferences and manifest it in the generated output. This task necessitates the integration of both modalities: relying solely on images would cause the model to blindly conflate visual features, while relying solely on text would leave the stylistic preference undefined, causing the model to collapse into its pre-trained distribution.

Ad-hoc Constraint Execution: A significant ability of *FI* is to perform dynamic reasoning under newly defined, ad-hoc rules. To systematically assess this, we construct two complementary tasks: (I) *Visual Constraint Generation* (II) *Symbolic Constraint Generation*, where novel meanings are assigned to symbols or images within the context. Crucially, we deliberately select elements that are devoid of pre-existing semantic associations, such as an image patch (e.g., "defining a plain blue square as an operation to remove the specifc object") or a mathematical symbol (e.g., "defining a function f as an instruction to make the object melt"). This design reflects the model's capability to solve novel problems by reasoning abstractly. The absence of either visual or textual modalities would fundamentally compromise the establishment of these ad-hoc rules, making it impossible to link the element to its new definition.

Contextual Knowledge Adaptation: A model possessing *FI* must exhibit flexible adaptation rather than rigid adherence to pretrained knowledge. We introduce two novel tasks to evaluate this: (I) *Prior-Conflicting Generation:* The model must reason based on newly defined common sense presented in the context (e.g., "object weight is determined by color"), even if it contradicts established facts. (II) *Multi-Semantic Generation:* This task requires the model to discern whether to interpret a concept literally or metaphorically (e.g., distinguishing "a green hand" as a novice vs. green skin) based on the specific multimodal contexts. In both tasks, missing any modality prevents the precise definition of new knowledge or the clarification of semantics, causing the model to fail in dynamic adaptation.

2.3 Evaluation Metric

Evaluating the ability of *GFI* remains a challenging task. Recent studies (Gu et al., 2024) have shown that large multimodal models exhibit strong visual reasoning and alignment capabilities, making them suitable as evaluators. Following this, we construct a robust pipeline comprising three orthogonal metrics. We utilize the Gemini-3-Pro (Google DeepMind, 2025) and Qwen2.5-VL-72B Bai et al. (2025) as the judge, employing structured prompts (Detailed in the Appendix A.2) and a hybrid evaluation strategy that incorporates manually-curated hints to ensure quantification. All metrics are assigned a score of 0 (fail), 1 (partial), or 2 (perfect).

Rule Compliance This metric measures the model's accuracy in executing strict ad-hoc rules. Since *GFI* tasks involve precise constraints (e.g., specific symbolic, layouts or palette), relying on the LMM's unguided interpretation is unreliable. We therefore adopt a hybrid protocol that grounds automated evaluation in human-verified truth. For each sample, we provide a manually curated eval-hint serving as the "gold standard". The LMM evaluator strictly compares the output against the specific nouns, adjectives and spatial predicates defined in this eval-hint.

Visual Consistency For some *GFI* tasks, the visual identity of original objects must remain unchanged (e.g., specific characters or objects). This metric assesses the model's ability to preserve context during dynamic reasoning. We provide specific hybrid eval-rubrics for each sample that identify the

Table 2: **Main Results.** We evaluate models across dimensions. The **Overall** column represents weighted score across all tasks, calculated using a ratio of RC:VC:AQ = 6:3.5:0.5. Ours is implemented on Bagel. The best and second best performances are highlighted.

Method	Interleaved	Overall	Implicit Pattern Induction			Ad-hoc Constraint Execution						Contextual Knowledge Adaptation					
			Implicit Pattern			*Symbolic Constraint*			*Visual Constraint*			*Prior-Conflicting*			*Multi-Semantic*		
			RC	VC	AQ	RC	VC	AQ	RC	VC	AQ	RC	VC	AQ	RC	VC	AQ
Proprietary Models																	
Nano Banana Pro	✓	57.19	66.86	44.59	96.51	71.38	50.00	92.11	76.67	66.67	96.67	52.97	41.38	90.59	35.45	-	95.00
Nano Banana	✓	50.66	56.47	39.04	94.12	60.46	51.91	90.20	68.33	79.17	93.33	35.50	39.47	91.00	30.28	-	93.12
GPT-Image	✗	47.15	58.14	41.92	93.60	58.82	32.82	93.79	49.17	62.50	92.50	43.50	33.33	90.00	28.64	-	85.45
SeeDream 4.0	✗	21.26	12.05	0.70	96.39	21.57	3.44	84.64	40.00	4.17	76.67	30.69	10.34	82.67	30.73	-	80.00
SeeDream 4.5	✗	52.84	70.00	59.59	97.06	62.91	41.09	94.37	58.33	62.50	86.67	40.10	41.38	92.57	35.00	-	86.82
Agentic Generation Models																	
Mind-Brush	✓	27.45	21.18	16.16	78.66	45.67	12.58	71.04	17.02	12.85	56.49	30.17	11.38	72.65	31.79	-	69.98
Gen-Searcher	✓	33.73	33.91	38.19	79.67	39.07	27.69	74.09	27.23	34.56	63.55	31.80	16.28	77.75	30.61	-	68.68
Open-Source Models																	
Qwen-Image	✗	30.58	36.18	27.69	71.05	36.18	27.69	71.05	26.67	45.83	55.83	27.72	20.69	71.78	25.91	-	69.55
GLM-Image	✗	24.71	32.94	19.86	93.53	22.37	21.15	87.50	27.50	12.50	70.83	20.30	15.52	71.29	17.73	-	70.91
FLUX.2-dev	✗	34.39	34.30	27.70	88.95	35.76	31.01	87.09	39.17	50.00	59.17	25.25	30.17	84.16	29.82	-	79.82
NextStep-1	✗	10.44	10.74	0.40	25.12	11.33	2.54	21.67	21.50	4.20	29.17	15.49	7.55	28.71	12.80	-	20.28
Emu3.5-Image	✗	36.67	41.86	35.81	83.72	34.97	39.31	86.93	24.17	29.17	42.50	26.24	37.93	82.18	32.87	-	75.46
Omini-Gen2	✗	27.87	29.07	26.35	76.16	25.33	30.38	77.96	11.67	41.67	52.50	23.76	34.48	69.80	19.27	-	63.76
Bagel	✓	26.74	26.74	27.03	84.30	29.61	16.03	76.32	22.50	12.50	49.17	22.28	17.24	74.75	33.49	-	53.67
Ours	✓	32.92	39.54	44.92	66.71	36.54	26.73	67.11	30.45	35.11	47.84	23.67	36.75	57.78	34.22	-	52.75

key visual elements from the reference images. The evaluator verifies the stability of these elements in generated image, ensuring model does not discard critical context while following instructions.

Aesthetic Quality Generative outputs must maintain physical coherence even when processing novel or conflicting semantic inputs. We employ a specific prompt to evaluate aesthetic quality, focusing on anatomical logic, lighting, and the presence of AI artifacts (e.g., distorted limbs). This metric ensures that the model's adaptation to new rules does not come at the cost of basic visual realism.

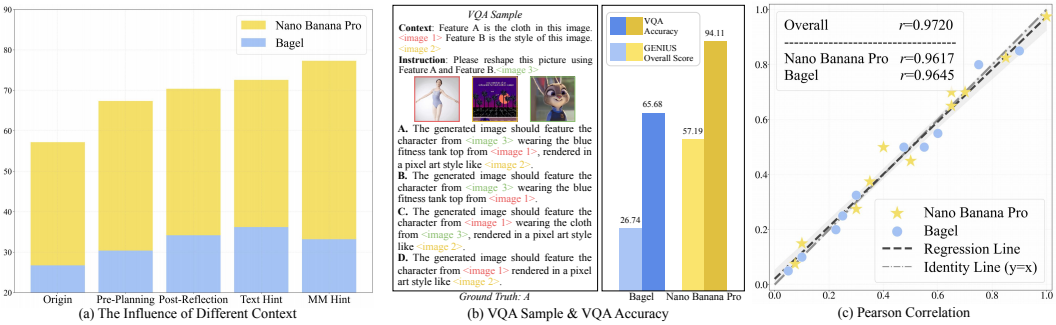
3 Experiment

We conduct a comprehensive evaluation of 14 representative open-source, proprietary and newly proposed agentic generation models. The open-source model comprises Qwen-Image-Edit-2511 (Wu et al., 2025a), GLM-Image (Zhipu AI Team, 2026), FLUX.2-dev (Labs, 2025), NextStep-1 (Team et al., 2025), Emu3.5-Image (Cui et al., 2025), Omini-Gen2 (Wu et al., 2026) and Bagel (Deng et al., 2025). The proprietary category includes leading commercial models: Nano Banana (Google, 2025a) and its Pro variant (Google, 2025b), SeeDream series (4.0 & 4.5) (Seedream et al., 2025) and GPT-Image (OpenAI Team, 2025). The agentic generation models includes Mind-Brush (He et al., 2026) and Gen-Searcher (Feng et al., 2026). Both of them use Qwen-Image as the backbone.

As outlined in Section 2.3, we employ Gemini-3-Pro (Google DeepMind, 2025) as the evaluator. To mitigate stochastic variance and ensure robustness, we report the final scores as the average of three independent runs for each sample. The quantitative results are shown in Table 2. Given the diversity of multimodal input formats, we adopt interleaved inputs for models capable of processing them, while utilizing a decoupled format for those that do not. Ablations on it are in Appendix D.1.

3.1 Main Results

***Generative Fluid Intelligence (GFI)* remains a significant bottleneck for current models.** Our results reveal a stark reality: even the state-of-the-art proprietary model, Nano Banana Pro, achieves an overall score of only 57.19. Meanwhile, representative open-source models like Bagel fall significantly behind, scoring a mere 26.74. Furthermore, recent newly proposed agentic generation model Gen-Searcher only gains +3.73% compared with its backbone Qwen-Image, while Mind-Brush even has a performance drop. These tasks demand ad-hoc reasoning and dynamic adaptation to novel rules, which challenge current models. Newly proposed agentic generation methods can not solve them either. Together, these quantitative deficits suggest that while current UMMs have acquired



(a) The Influence of Different Context (b) VQA Sample & VQA Accuracy (c) Pearson Correlation

Figure 2: **Diagnostic analysis and metric validation.** (a) Performance comparison across different context settings. (b) Analysis of the gap between context comprehension (VQA) and generation capabilities. (c) Correlation analysis validating the LMM-as-a-Judge metric.

robust capabilities for crystallized reproduction, they remain fundamentally distant from the fluid adaptability required for general-purpose generation.

Current models fail to effectively arbitrate the conflict between pre-trained priors and the given context. As shown in Table 2, this deficiency is most pronounced in the *Contextual Knowledge Adaptation* dimension, where performance consistently drops below other task categories. When ad-hoc instructions explicitly contradict world knowledge (e.g., counter-intuitive physical laws or remapped semantics), models exhibit a strong "cognitive inertia", frequently defaulting to their pre-trained priors. This suggests that existing architectures lack a robust mechanism to inhibit intrinsic priors, failing to dynamically adapt to the context.

High aesthetic fidelity can partially obscure underlying logical deficits. Our hybrid metric analysis uncovers a pervasive "illusion of competence": models consistently maintain high *Aesthetic Quality*, yet their performance on *Rule Compliance* lags substantially behind. This discrepancy suggests that previous model optimization has disproportionately focused on surface-level visual plausibility at the expense of deep context interpretation and logical adherence. By exposing this, **GENIUS** signals a necessary paradigm shift for next generation of models: moving beyond merely generating "beautiful" pixels to achieving profound context comprehension and ensuring logically correct visual synthesis.

3.2 Discussion and Analysis

Pre-planing and post-reflection yield marginal gains. We investigated various inference-time enhancement strategies to potentially mitigate performance deficits. Taking Nano Banana Pro and Bagel as examples, we implemented pre-planning (activating reasoning mode) and post-reflection (an iterative process where initial generations are evaluated and re-fed as context for refinement). However, as shown in Figure 2 (a), empirical results across both models indicate that these strategies yield only marginal gains (Compared with the improvements on other benchmarks like WISE). This suggests that current architectures struggle to effectively leverage explicit reasoning for generation.

Context comprehension is the key to solve *GFI* problems. To isolate the source of failure, we introduced human-curated hints to guide the generation process. Specifically, we employed a progressive intervention strategy: initially utilizing text-only hints, and subsequently constructing multimodal hints to ensure information completeness, thereby explicitly guiding the model's generation. The results are illustrated in Figure 2 (a). This intervention resulted in substantial performance improvements; however, the degree of improvements varied significantly: Nano Banana Pro exhibited a much more boost compared to Bagel. This observation highlights that accurate context comprehension acts as a critical factor in solving *GFI* tasks. Meanwhile, solving *GFI* problems requires not only accurate context comprehension to decode ad-hoc rules but also robust intrinsic model capabilities to execute them, implying that comprehension aids cannot fully compensate for a weaker backbone.

Recognizing the intended outcome does not ensure faithful generation. To investigate the root cause, we reformulated the generative tasks into comprehension-oriented Visual Question Answering probes. Specifically, we structured these probes as multiple-choice questions that query the model regarding the expected visual appearance of the target image. We rewrite the human-authored RC evaluation hints as the correct answer option and construct three distractors for each instance, while

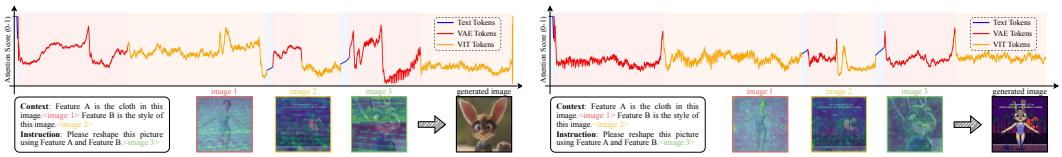
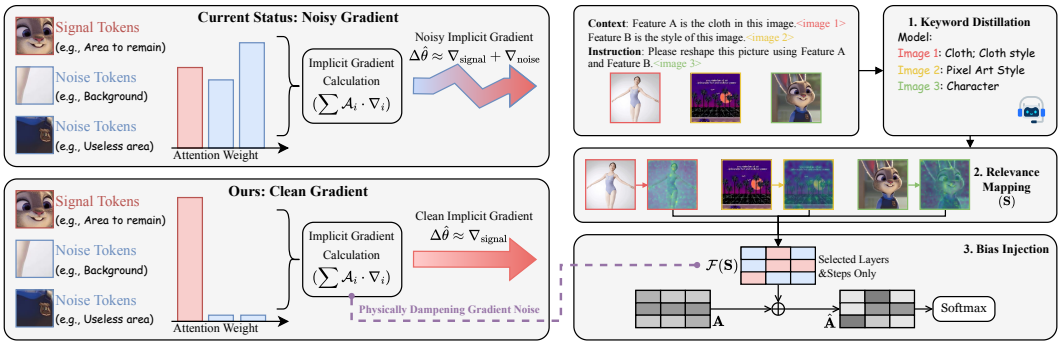


Figure 3: **Visualization of attention scores (range [0, 1]).** Left: Existing models. Right: Ours.



(a) Theoretical Impact of Attention on Implicit Gradients (b) The Method Pipeline

Figure 4: **Method.** Guided by the theoretical insight that attention magnitude dictates implicit gradient norms (a), we implement a three-stage pipeline (b) to explicitly suppress noise tokens and rectify the implicit optimization direction.

simultaneously constructing three distractors for each sample to facilitate evaluation. The results are shown in Figure 2 (b). Empirical results reveal a significant disparity: models frequently demonstrate an accurate understanding of the context's intent but fail to translate this into compliant visual outputs. This suggests that the model's current cognitive processing of the context, while sufficient for discriminative understanding tasks, lacks the granularity required for generative reconstruction. We hypothesize this stems from two factors: first, the high information density of interleaved contexts, where fine-grained visual nuances (e.g., specific textures) are difficult to fully capture and articulate through limited modalities; and second, a structural inefficiency in current UMM architectures, where rich semantic understanding from the encoder is not effectively propagated to the generative decoder, resulting in a "know-but-cannot-draw" phenomenon. We further discuss it in Section 4.1.

Together, the VQA and hint-based analyses examine complementary aspects of context-conditioned generation. The VQA probe tests recognition of the intended outcome, while the hint interventions assess generation with explicit target guidance. Their results motivate examining both the extraction of contextual requirements and their execution during image synthesis.

3.3 Validity of LMM-as-a-Judge

To assess the reliability of LMM-based evaluation, we compared automatic scores with human judgments on output images from Nano Banana Pro and Bagel across all dimensions. Five human experts independently rated the samples using the same metrics.

As shown in Figure 2 (c), LMM scores closely align with human ratings, with Pearson correlations of $r = 0.9617$ for Nano Banana Pro and $r = 0.9645$ for Bagel. Dimension-specific analysis further yields Mean Absolute Errors (MAE) of 0.06–0.11 across metrics, small relative to the 0–2 scoring scale. These results demonstrate that the evaluator captures human judgments consistently across models and task dimensions, supporting LMM-as-a-Judge as a reliable alternative to human.

To further assess reproducibility and cross-model robustness, we also employed the open-source Qwen2.5-VL-72B (Bai et al., 2025) as a judge. As reported in Appendix C, Qwen2.5-VL-72B assigns systematically lower absolute scores than Gemini-3-Pro, suggesting stricter evaluation criteria, while preserving broadly consistent relative performance trends and largely similar model rankings. This agreement shows that the main comparative findings remain consistent despite differences in the evaluators' scoring strictness. Consistency across proprietary and open-source judges confirms that the observed performance gaps reflect the evaluated models rather than artifacts of a particular judge, reinforcing the reliability and generalizability of the results.

4 A POTENTIAL SOLUTION

Evaluation on **GENIUS** reveals a clear gap between current SOTA models and general intelligence. To diagnose the potential causes of this deficit, we conduct a comprehensive analysis from both theoretical and empirical perspectives, focusing on the widely applicable Bagel framework.

4.1 EXPERIMENTAL OBSERVATION

To investigate the underlying mechanism of failure, we visualized the attention distribution over the entire context, using the image tokens generated during the process as the query. Surprisingly, as shown in the left part of Figure 3, we found the attention distribution is unreasonable: it exhibits irregular noise and stochastic spikes across the multimodal context. This indicates the model struggles to precisely capture pivotal ad-hoc rules from the context. Instead of pinpointing the critical definition, the attention is spread out indiscriminately across the input. As a result, the model fails to extract the specific signal needed for adaptation and simply falls back to its pre-trained priors.

4.2 THEORETICAL ANALYSIS

To explain this phenomenon, we adopt the theoretical perspective of *In-Context Learning (ICL) as Implicit Fine-Tuning* from (Dherin et al., 2025; Ahn et al., 2023; Dai et al., 2023; von Oswald et al., 2023). we perform a derivation on Bagel, which adopts a Mixture-of-Transformer architecture. Since **GENIUS** targets the generative task, we redefine the function $\mathcal{A}(u, g)$ from (Dherin et al., 2025), based on the Bagel model, where $\mathcal{A}$ denotes the network layer component responsible for context processing, u represents the encoding of context and instructions, and g denotes the encoding of intermediate noisy tokens of images. We suppose in the t-th step and the l-th Decoder blocks we have $g^{(t,l+1)} = \mathcal{L}^{(l)}_{\mathrm{Up},b}(u^{(l)}, g^{(t,l)})$, where Up is a projection layer in the decoder block, b is the bias of Down layer, and $\mathcal{L}$ represents the l-th block's forward propagation. And then we can formalize the relationship between u and the (Up, b):

Theorem 4.1. *The layer update satisfies following property:*

$$\mathcal{L}_{Up+\Delta Up,\, b+\Delta b}(u', g) = \mathcal{L}_{Up,b}(u, g),\ \Delta b = \mathcal{A}(u, g) - \mathcal{A}(u', g) \tag{1}$$

where the upsampling operator perturbation be defined as:

$$\Delta Up = \frac{Up\,(\delta\mathcal{A})\,\mathcal{N}(\mathcal{A}(u', g))^\top}{\|\mathcal{N}(\mathcal{A}(u', g))\|^2},\ \mathcal{N}(x) = \frac{x}{RMS(x)} \tag{2}$$

And the normalized attention difference is given by:

$$\delta\mathcal{A} = \mathcal{N}(\mathcal{A}(u, g)) - \mathcal{N}(\mathcal{A}(u', g)) \tag{3}$$

According to Theorem 4.1, the ICL process is mathematically equivalent to updating specific model parameters. This successfully extends the conclusions of (Dherin et al., 2025):

We formalize the vector notations: let $u = (u_1, \ldots, u_n)$, and $u^{(i)} = (u_1, \ldots, u_i)$. Then we hope:

$$\mathcal{L}_{\mathrm{Up}_i, b_i}(u^{(i)}, g) = \mathcal{L}_{\mathrm{Up},b}(u, g) \quad \forall i = 1, 2, \ldots, n \tag{4}$$

From this objective, we derive expressions for Up_i and b_i:

$$\mathrm{Up}_i = \mathrm{Up} + \Delta\mathrm{Up}_i = \mathrm{Up} + \frac{\mathrm{Up}\,(\delta\mathcal{A}_i)\,\mathcal{N}(\mathcal{A}(g))^\top}{\|\mathcal{N}(\mathcal{A}(g))\|^2},\ b_i = b + \Delta b_i = b + \mathcal{A}(u^{(i)}, g) - \mathcal{A}(g) \tag{5}$$

where the attention difference term $\delta\mathcal{A}_i$ is defined as:

$$\delta\mathcal{A}_i = \mathcal{A}(u^{(i)}, g) - \mathcal{A}(g). \tag{6}$$

Based on the above derivations, we present the key theorem for iterative parameter updates:

Theorem 4.2. *For the $(i+1)$-th iteration, Up and b follow the gradient descent update rules below:*

$$\begin{cases} Up_{i+1} = Up_i - h\nabla_{Up} L_i(Up_i), \\ b_{i+1} = b_i - \nabla_b \left(\mathrm{tr}\left(\delta_i^\top b_i\right)\right) \end{cases} \tag{7}$$

where $h = 1/\left\|\mathcal{N}(\mathcal{A}(g))\right\|^2$ denotes the learning rate. $L_i(\text{Up}) = \text{tr}\left(\Delta_i^\top \text{Up}\right)$ is loss function, among which $\Delta_i = \text{Up}\left(\hat{\delta}_i\right)\mathcal{N}(\mathcal{A}(g))^\top$, $\hat{\delta}_i = \mathcal{N}(\mathcal{A}(u^{(i)}, g)) - \mathcal{N}(\mathcal{A}(u^{(i+1)}, g))$, and $\delta_i = \mathcal{A}(u^{(i)}, g) - \mathcal{A}(u^{(i+1)}, g)$. Combining empirical observations with theoretical analysis, we offer a hypothesis for the deficit in *GFI*: The imbalanced attention distribution results in a lack of guidance during implicit gradient descent. Consequently, the descent direction becomes stochastic, failing to overcome the pre-trained priors. Full proof of both theorems is in the Appendix G.

4.3 Attention Adjustment Mechanism

Guided by Theorem 4.2, we recognize that the magnitude of attention assigned to the context directly dictates the norm of the implicit gradient update. The irregular attention distribution within the context images, as previously observed in Section 4.1, implies that irrelevant "noise" tokens currently contribute significant, erroneous gradient components, thereby diverting the optimization trajectory away from the optimal path. To counteract this, we propose a training-free adjustment mechanism to recalibrate the update direction. By explicitly suppressing the attention weights of noise tokens, we mathematically dampen their corresponding gradient norms (i.e., $\|\Delta \text{Up}_{\text{noise}}\| \to 0$), ensuring the implicit fine-tuning is driven solely by critical context signals.

Specifically, we implement this mechanism through a three-stage pipeline shown in Figure 4 (b). First, in the Keyword Distillation phase, leveraging the semantic reasoning capability of Bagel, we prompt the model to distill task-critical visual cues into a set of region-specific keywords $\mathcal{K}$ (the prompt is detailed in the Appendix F.1). Subsequently, during Relevance Mapping, we compute a semantic relevance map $\mathbf{S}$ by evaluating the alignment between these keywords and the visual context tokens, where $\mathbf{S}$ serves as a proxy for the token's contribution to the effective gradient signal. Finally, via Bias Injection, we inject a spatial bias $\mathcal{F}(\mathbf{S})$ directly into the attention logits:

$$\text{Attention} = \textit{Softmax}\left(\frac{\mathbf{A} + \lambda \cdot \mathcal{F}(\mathbf{S})}{\sqrt{d}}\right) V \tag{8}$$

This formulation ensures tokens with high relevance are emphasized while noise is suppressed. The detailed mathematical formulation is provided in the Appendix F.2. By rectifying the attention landscape, we re-weight the implicit gradient updates, deterministically steering the optimization trajectory to overcome pre-trained priors.

4.4 Experimental Results

As visualized in Figure 3 (Right), our mechanism successfully rectifies the originally disordered attention landscape into a sharpened distribution with distinct peaks focused on critical tokens. Quantitative results in Table 2 demonstrate consistent performance gains across nearly all dimensions compared to Bagel. This validates that deterministically steering the implicit gradient trajectory effectively activates the model's latent *GFI* without requiring parameter updates, establishing a simple yet strong baseline. Moreover, we provide additional generalization experiments in the Appendix D.3.

5 Conclusion

We introduced **GENIUS**, the first benchmark to systematically quantify *Generative Fluid Intelligence (GFI)*. Grounded in Cattell-Horn-Carroll (CHC) theory, we formalized *GFI* into three core dimensions: *Implicit Pattern Induction*, *Ad-hoc Constraint Execution*, and *Contextual Knowledge Adaptation*, establishing a rigorous standard for evaluating models in novel, reasoning-intensive scenarios. Evaluating 14 representative open-source and proprietary models reveals that even state-of-the-art models like Nano Banana Pro achieves an overall score of only 57.19, while open-source models exhibit significant deficits. Our analysis exposes an "execution gap": models struggle to resolve conflicts between pre-trained priors and ad-hoc context, often prioritizing aesthetic fidelity over logical rule compliance. We partially attribute these failures to inference-time attention defects and propose a training-free adjustment strategy that effectively activates latent *GFI* capabilities. We hope **GENIUS** serves as a pivotal testbed for future research, guiding next-generation models from crystallized memorization toward true general intelligence.

AI USE STATEMENT

LMMs are used as components of our evaluation and intervention pipelines. Gemini-3-Pro serves as the primary evaluator, and Qwen2.5-VL-72B provides an alternative evaluation using the same human-authored RC/VC hints and shared AQ rubric. LLM assistants also supported manuscript revision and wording. The authors take responsibility for the research, results, and final manuscript.

REPRODUCIBILITY STATEMENT

We describe the benchmark construction, evaluation metrics, tested models, and input-format settings in the main text. GENIUS contains 510 expert-curated instances, and reported scores are averaged over three independent generations per instance. Appendix A.2 provides the evaluation prompts, while Appendix B illustrates inputs, evaluation hints, and model outputs. Appendix C reports results with an alternative judge, and Appendices D.1 and D.2 examine input formatting and context removal. The keyword-extraction prompt, attention-modulation formulation, and intervention settings are provided in Appendices F.1, F.2, and F.3, respectively.

ETHICS STATEMENT

GENIUS evaluates visual generation under context-defined rules. Human involvement consists of expert task curation, annotation, and output assessment, including independent ratings by five experts. Prior-conflicting examples specify hypothetical scenarios for evaluating contextual adaptation. We do not anticipate significant ethical concerns specific to this study. The benchmark is intended to support research on reliable, context-sensitive image generation.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning, 2023. URL https://arxiv.org/abs/2306.00297.

Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, et al. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*, 2025.

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

Tomer Barak and Yonatan Loewenstein. Investigating learning-independent abstract reasoning in artificial neural networks. *arXiv e-prints*, pp. arXiv–2407, 2024.

Raymond B Cattell. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1, 1963.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Wei Chow, Jiachun Pan, Yongyuan Liang, Mingze Zhou, Xue Song, Liyu Jia, Saining Zhang, Siliang Tang, Juncheng Li, Fengda Zhang, et al. Weave: Unleashing and benchmarking the in-context interleaved comprehension and generation. *arXiv preprint arXiv:2511.11434*, 2025.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners, 2025. URL https://arxiv.org/abs/2510.26583.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers, 2023. URL https://arxiv.org/abs/2212.10559.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, and Javier Gonzalvo. Learning without training: The implicit dynamics of in-context learning, 2025. URL https://arxiv.org/abs/2507.16003.

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

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*, 2026.

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

Google. Introducing Gemini 2.5 Flash Image, our state-of-the-art image model. https://developers.googleblog.com/introducing-gemini-2-5-flash-image/, 2025a.

Google. Introducing Nano Banana Pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025b.

Google DeepMind. Gemini 3 Pro. https://deepmind.google/models/gemini/pro/, 2025.

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

Ziyu Guo, Renrui Zhang, Hongyu Li, Manyuan Zhang, Xinyan Chen, Sifan Wang, Yan Feng, Peng Pei, and Pheng-Ann Heng. Thinking-while-generating: Interleaving textual reasoning throughout visual generation. *arXiv preprint arXiv:2511.16671*, 2025.

Ziyu Guo*, Renrui Zhang*, Chengzhuo Tong*, Zhizheng Zhao*, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *CVPR 2025*, 2025.

Jun He, Junyan Ye, Zilong Huang, Dongzhi Jiang, Chenjue Zhang, Leqi Zhu, Renrui Zhang, Xiang Zhang, and Weijia Li. Mind-brush: Integrating agentic cognitive search and reasoning into image generation. *arXiv preprint arXiv:2602.01756*, 2026.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Susanne M Jaeggi, Martin Buschkuehl, John Jonides, and Walter J Perrig. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, 105(19): 6829–6833, 2008.

Dongzhi Jiang*, Ziyu Guo*, Renrui Zhang*, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

Weiyang Jin, Yuwei Niu, Jiaqi Liao, Chengqi Duan, Aoxue Li, Shenghua Gao, and Xihui Liu. Srum: Fine-grained self-rewarding for unified multimodal models. *arXiv preprint arXiv:2510.12784*, 2025.

Phillip Kent. Fluid intelligence: A brief history. *Applied Neuropsychology: Child*, 6(3):193–203, 2017.

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025a.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 415–423, 2025b.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Libo Qin, Ruihan Chen, Baohang Li, Kui Jiang, Yaowei Wang, Ting Liu, et al. Cai: Caption-sensitive attention intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2506.23590*, 2025c.

Yanshu Li, JianJiang Yang, Bozheng Li, and Ruixiang Tang. Cama: Enhancing multimodal in-context learning with context-aware modulated attention. *arXiv preprint arXiv:2505.17097*, 2025d.

Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. Unieval: Unified holistic evaluation for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.10483*, 2025e.

Yongyuan Liang, Wei Chow, Feng Li, Ziqiao Ma, Xiyao Wang, Jiageng Mao, Jiuhai Chen, Jiatao Gu, Yue Wang, and Furong Huang. Rover: Benchmarking reciprocal cross-modal reasoning for omnimodal generation. *arXiv preprint arXiv:2511.01163*, 2025.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

OpenAI Team. New ChatGPT Images is Here. https://openai.com/index/new-chatgpt-images-is-here/, 2025.

Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. Diffusiongpt: Llm-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*, 2024.

Stefan Schipolowski, Oliver Wilhelm, and Ulrich Schroeders. On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence*, 46:156–168, 2014.

W Joel Schneider and Kevin S McGrew. The cattell-horn-carroll model of intelligence. 2012.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, et al. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. *arXiv preprint arXiv:2509.24897*, 2025.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Ailin Huang, Bin Wang, Changxin Miao, Deshan Sun, En Yu, Fukun Yin, Gang Yu, Hao Nie, Haoran Lv, Hanpeng Hu, Jia Wang, Jian Zhou, Jianjian Sun, Kaijun Tan, Kang An, Kangheng Lin, Liang Zhao, Mei Chen, Peng Xing, Rui Wang, Shiyu Liu, Shutao Xia, Tianhao You, Wei Ji, Xianfang Zeng, Xin Han, Xuelin Zhang, Yana Wei, Yanming Xu, Yimin Jiang, Yingming Wang, Yu Zhou, Yucheng Han, Ziyang Meng, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, and Yibo Zhu. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent, 2023. URL https://arxiv.org/abs/2212.07677.

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

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025b.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025a.

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

Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models. *arXiv preprint arXiv:2504.03641*, 2025b.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9050–9061, 2025.

Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K Wong. Bridging different language models and generative vision models for text-to-image generation. In *European Conference on Computer Vision*, pp. 70–86. Springer, 2024.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

Zhipu AI Team. https://z.ai/blog/glm-image, January 2026.

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

A Benchmark Details

A.1 Data Statistics

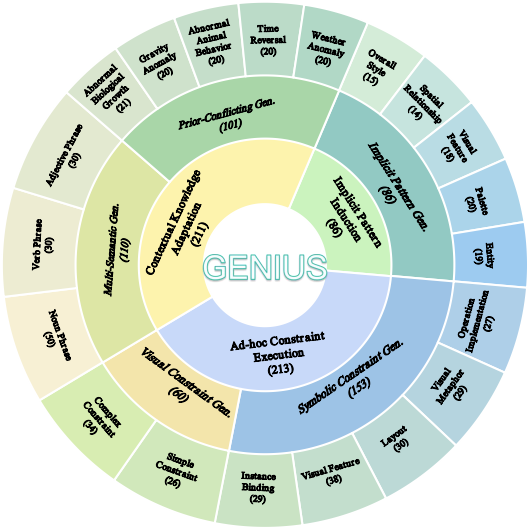


Figure 5: **Data composition pie chart. GENIUS** comprises 3 dimensions, 5 tasks, and 20 sub-tasks.

A.2 Evaluation Prompt

As illustrated in the previously presented prompt templates, we developed a evaluation framework using Large Multimodal Models (LMMs) to assess three key dimensions of generative quality:

Rule Compliance (RC): For each **GENIUS** sample, an audit of textual-visual alignment is conducted. This process rigorously verifies nouns, adjectives, and spatial constraints to ensure 100% compliance with specific modification requests. Details of the prompt template are provided in Figure 8.

Visual Consistency (VC): For each **GENIUS** sample, *Visual Consistency* may be evaluated multiple times or not at all, depending on how many reference images (objects in the image) need to remain visually consistent. Since we have observed that many open-source models directly copy reference images to cheat (e.g., Bagel, GLM-Image, etc.), a dedicated anti-plagiarism screening is conducted prior to the *Visual Consistency* audit. The LMM first performs a pixel-level identity check; if the target image is found to be an exact pixel-for-pixel duplicate of the reference without any generative modifications, the consistency score is automatically set to 0. This is because consistency must be built on editing rather than plagiarism. Details of the prompt template are provided in Figure 9.

Aesthetic Quality (AQ): Assesses visual logic, rendering clarity, and realism. It rewards commercial-grade outputs while penalizing structural collapses or AI hallucinations. Details of the prompt template are provided in Figure 10.

B Detailed Qualitative Examples and Model Outputs

To provide a more granular view of the **GENIUS** benchmark, we present comprehensive qualitative examples for each sub-task. For every sample, we showcase a complete instance that includes: (1) the full input content (comprising both context and instruction); (2) the specific evaluation hints utilized for assessing *Rule Compliance (RC)* and *Visual Consistency (VC)*; and (3) the corresponding generated outputs from six representative models: Nano Banana Pro Google (2025b), Nano Banana Google (2025a), SeeDream4.5 Seedream et al. (2025), FLUX.2-dev Labs (2025), Bagel Deng et al. (2025) and ours. These detailed comparisons, which highlight the capabilities and failure modes of different architectures, are illustrated in Figure 11 and Figure 12.

Table 3: **Benchmark Results by Qwen2.5-VL-72B.** The **Overall** column represents the weighted score across all tasks, calculated using a metric ratio of RC:VC:AQ = 6:3.5:0.5. The best and second best performances are highlighted.

Method	Interleaved	Overall	Implicit Pattern Induction			Ad-hoc Constraint Execution						Contextual Knowledge Adaptation					
			Implicit Pattern			*Symbolic Constraint*			*Visual Constraint*			*Prior-Conflicting*			*Multi-Semantic*		
			RC	VC	AQ	RC	VC	AQ	RC	VC	AQ	RC	VC	AQ	RC	VC	AQ
Proprietary Models																	
Nano Banana Pro	✓	48.35	62.21	37.84	89.53	69.93	34.35	82.68	74.14	54.17	88.33	28.22	42.24	89.11	27.27	-	67.27
Nano Banana	✓	42.88	52.72	41.89	80.81	58.52	35.50	80.72	66.95	45.83	89.83	22.24	28.45	85.64	23.00	-	68.64
GPT-Image	✗	40.94	53.15	41.27	90.35	59.15	29.77	82.35	42.50	33.33	69.17	27.72	35.34	89.60	17.73	-	58.18
SeeDream 4.0	✗	17.74	8.72	1.35	92.44	19.93	5.73	76.47	37.50	8.33	76.67	11.39	8.62	85.64	22.82	-	68.18
SeeDream 4.5	✗	44.17	64.79	40.32	89.65	62.11	25.08	80.39	60.83	39.50	89.17	26.80	40.66	84.16	26.18	-	62.27
Agentic Generation Models																	
Mind-Brush	✓	22.52	17.45	10.06	88.18	43.33	11.2	74.17	13.29	7.01	59.87	13.69	16.59	79.66	19.41	-	68.6
Gen-Searcher	✓	26.82	27.94	23.78	89.31	33.83	24.66	77.35	21.27	18.85	67.35	14.43	23.74	85.26	18.69	-	67.33
Native Unifed Models																	
SenseNova-U1-8B	✓	28.32	34.88	28.03	86.60	38.24	19.70	81.08	32.50	29.17	78.33	6.44	20.81	84.58	21.91	-	74.55
Open-Source Models																	
Qwen-Image	✗	25.67	29.81	17.24	79.65	31.33	24.66	74.18	20.83	25.00	59.17	12.58	30.17	78.71	15.82	-	68.18
GLM-Image	✗	17.45	23.23	17.22	80.23	15.99	21.81	71.76	23.33	18.67	69.17	6.44	15.93	79.21	10.09	-	48.64
FLUX.2-dev	✗	27.37	33.40	17.57	85.47	33.32	27.36	77.06	30.33	37.70	63.75	12.87	30.38	80.69	19.27	-	63.18
NextStep-1	✗	9.90	0.38	20.02	11.98	1.56	15.22	20.54	3.19	19.32	14.11	6.98	20.21	10.08	12.28	-	13.57
Emu3.5-Image	✗	28.80	43.02	26.35	80.81	33.66	26.72	81.70	20.83	37.50	45.00	10.89	24.14	84.65	20.45	-	62.73
Omini-Gen2	✗	21.12	24.42	18.24	81.40	20.59	22.90	82.03	8.33	20.83	62.50	11.39	31.90	82.67	8.18	-	58.64
Bagel	✓	18.97	14.53	20.27	80.23	16.01	14.89	80.72	16.67	25.00	57.50	6.93	16.38	82.67	22.73	-	60.45
Ours	✓	23.91	26.45	30.71	70.54	22.01	20.24	75.53	25.27	26.61	46.37	7.89	27.55	75.92	22.35	-	47.24

C Evaluation using open source model.

To further validate the robustness of our evaluation framework, we employed Qwen2.5-VL-72B Bai et al. (2025) as the judge model to assess **GENIUS** benchmark. The results are summarized in Table 3.

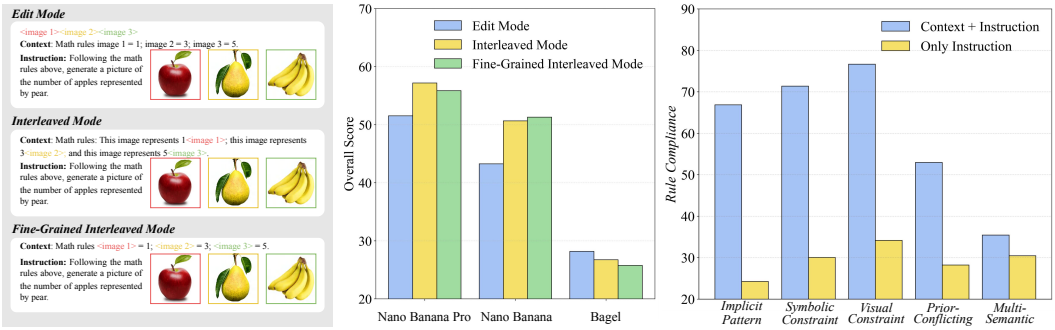
As illustrated, utilizing Qwen2.5-VL-72B as the evaluator results in a universal decrease in the Overall Scores across all tested models. This suggests that Qwen2.5-VL-72B may impose a stricter standard for rule and visual compliance compared to the primary evaluator. Crucially, despite the shift in absolute scores, the relative performance trends and the ranking order of the models remain largely consistent. This consistency reinforces the reliability of **GENIUS**, showing that the observed performance gaps are intrinsic to the models themselves rather than artifacts of a specific judge.

D Additional Experiments and Analysis

D.1 Ablation on Interleaved Format

In the context of the **GENIUS** Benchmark, multimodal interleaved data can be presented in various input formats. Since models exhibit varying degrees of compatibility with these formats, we investigate the impact of input structure on performance by defining three distinct paradigms, as illustrated in Figure 6(a). First, in the Edit Mode, the visual and textual modalities are decoupled. Images are provided separately (e.g., appended at the end or beginning) and are referenced within the text using placeholders like "image i". Second, the Interleaved Mode corresponds to the standard setting used in our main experiments. Here, images are interleaved with text but are inserted at the boundaries of complete semantic units (typically at the end of a sentence), preserving the syntactic integrity of the text strings. Third, the Fine-Grained Interleaved Mode inserts images precisely at their point of reference, even within a sentence. In this mode, visual tokens act as intrinsic parts of the syntax and can interrupt the textual flow, requiring the model to handle fine-grained multimodal dependencies.

We conducted evaluations on the Nano Banana series and Bagel, as they are among the few models capable of supporting all three formats. The Overall scores are reported in Figure 6(b). The results indicate that performance trends vary across models, likely due to differences in model architecture. Notably, we observe a significant performance gap between Edit Mode and the two interleaved modes (Interleaved and Fine-Grained), while the disparity between the two interleaved formats is relatively



(a) Different Interleaved Format (b) Overall Scores for Different Interleaved Formats (c) *Rule Compliance* for Different Input Composition

Figure 6: **Additional Experiments and Analysis.** This figure presents the definition of three input formats (a) and their corresponding performance impact (b), followed by an ablation study assessing the importance of contextual information in instruction following (c).

marginal. It suggests that current multimodal models possess limited robustness regarding input formatting, exhibiting a strong sensitivity to how visual information is integrated with text.

D.2 Discussion on the Composition of Input

To verify the necessity of contextual information for high-fidelity generation, we conducted an ablation study on the Nano Banana Pro model by removing the context component and relying solely on the final instruction. The comparative *Rule Compliance* scores across different tasks are reported in Figure 6(c). As observed, removing context leads to a precipitous decline in performance across the board, underscoring its indispensable role. Specifically, the *Implicit Pattern Generation*, *Symbolic Constraint Generation*, and *Visual Constraint Generation* tasks suffer the most severe degradation. This is anticipated, as these tasks require the model to inductively reason or extract specific visual-textual mappings defined solely within the context; without these definitions, the model lacks the necessary premises to execute the instruction. Similarly, the *Prior-Conflicting Generation* task exhibits a significant drop, as the model inevitably reverts to its pre-trained priors in the absence of an explicit counter-factual context to override them. Interestingly, the decline in the *Multi-Semantic Generation* task is less pronounced. This relative stability can be attributed to the task's inherent difficulty (resulting in a lower baseline performance) and the probability that the model might fortuitously align with the target semantics even without disambiguating context. Nevertheless, the consistent performance gap confirms that context is not merely supplementary but a critical foundation for accurate generation in complex scenarios.

D.3 Generalization to Other Models

To further examine the generalizability of our proposed attention modulation strategy, we apply it to GLM-Image (Zhipu AI Team, 2026). Since our method operates by modulating attention logits according to keyword-based relevance scores, it does not rely on architecture-specific training or model-specific parameters. Therefore, it can be naturally transferred to other generation models with accessible attention modules. As shown in Table 4, applying our attention modulation strategy to GLM-Image brings consistent quantitative improvements, increasing the overall score from 24.71 to 30.04. These results demonstrate that our method is not limited to a single backbone, but provides a robust and architecture-agnostic mechanism for enhancing generative follow-through ability.

E Related Work

Fluid Intelligence Originating from the Cattell-Horn-Carroll (CHC) theory of cognitive abilities Schneider & McGrew (2012), general intelligence is structurally divided into *Crystallized Intelligence* (G_c) and *Fluid Intelligence* (G_f). While G_c relies on the utilization of accumulated knowledge, G_f represents the innate capacity to solve novel problems through inductive reasoning and dynamic reasoning, independent of prior knowledge, which is often considered as more indicative

Table 4: **Generalization results on GLM-Image.** The Overall column reports the weighted score across all tasks, computed with the ratio RC:VC:AQ $= 6 : 3.5 : 0.5$. "Ours" denotes applying our proposed intervention strategy on top of GLM-Image.

Method	Interleaved	Overall	Implicit Pattern Induction			Ad-hoc Constraint Execution						Contextual Knowledge Adaptation					
			Implicit Pattern			*Symbolic Constraint*			*Visual Constraint*			*Prior-Conflicting*			*Multi-Semantic*		
			RC	VC	AQ	RC	VC	AQ	RC	VC	AQ	RC	VC	AQ	RC	VC	AQ
GLM-Image	✗	24.71	32.94	19.86	93.53	22.37	21.15	87.50	27.50	12.50	70.83	20.30	15.52	71.29	17.73	-	70.91
Ours	✗	30.04	39.87	37.18	82.78	26.01	27.44	81.23	30.67	32.50	60.33	20.89	33.27	63.23	17.80	-	65.62

of general intelligence Jaeggi et al. (2008); Chollet (2019); Barak & Loewenstein (2024). In the field of understanding, evaluating G_f has traditionally focused on logical reasoning and abstract pattern completion. Prominent benchmarks such as the ARC Bench Chollet (2019) assess a model's ability to induce rules from few-shot examples and generalize to new scenarios. However, these evaluations are predominantly discriminative or symbolic, targeting problem-solving in restricted domains (e.g., grid worlds). In the context of Unified Multimodal Models (UMMs), current assessments remain largely confined to G_c, testing the model's capability on static world knowledge.

Unified Multimodal Models (UMMs) Recent years have witnessed a paradigm shift from modular composition towards native fusion in multimodal models. Early approaches primarily bridged pre-trained Large Language Models Qin et al. (2024); Esser et al. (2024); Zhao et al. (2024); Li et al. (2025b); Zhang et al. (2025) with diffusion decoders to enable visual synthesis capabilities Koh et al. (2023). However, the latest wave of UMMs, represented by Chameleon Team (2024), Show-o Xie et al. (2024); Guo* et al. (2025) and Emu Series Sun et al. (2023; 2024); Wang et al. (2024), marks a fundamental departure by discretizing visual signals into discrete tokens. Janus Wu et al. (2025b) and its improvements Guo et al. (2025); Jiang* et al. (2025) claims that understanding and generation require distinct information, employing different tokenizers for each task. Among these architectures, Bagel Deng et al. (2025) and its improvements Xie et al. (2025a); Jin et al. (2025) have demonstrated remarkable versatility in both understanding and generation tasks, becoming one of the most representative open-sourced models. Despite these architectural breakthroughs, current research predominantly evaluates these models on task-specific benchmarks (e.g., VQA or standard T2I generation, *Crystallized Intelligence*), leaving their intrinsic *Generative Fluid Intelligence* (i.e., the capacity to reason and adaptively generate under novel, ad-hoc constraints) largely unexplored.

Generative Evaluation of UMMs With the rapid progress of UMMs, numerous benchmarks Ghosh et al. (2023); Zhao et al. (2025); Wei et al. (2025); Li et al. (2025e); Chow et al. (2025) have been proposed to assess their capabilities, yet most remain confined to traditional evaluation paradigms. Early benchmarks like GenEval Ghosh et al. (2023), WISE Niu et al. (2025), and DPG-Bench Hu et al. (2024) primarily focus on single-image generation tasks, assessing static world knowledge or basic text-to-image alignment without involving complex, interleaved contexts. OpenIng Zhou et al. (2025) focuses on in-context visual generation, while it primarily targeting interleaved text-and-image generation and *Crystallized Intelligence*. While more recent suites such as MME-Unify Xie et al. (2025b), RealUnify Shi et al. (2025), and ROVER Liang et al. (2025) have begun to incorporate multi-image inputs, they predominantly target *Crystallized Intelligence*, evaluating the model's ability to recall pre-trained information rather than adapting novel rules. Crucially, none of the existing benchmarks systematically evaluate *Generative Fluid Intelligence (GFI)*. As shown in Table 1, current methods lack comprehensive coverage across key GFI dimensions, including Implicit Pattern Induction, Explicit Constraint Execution, and Contextual Knowledge Adaptation. Furthermore, many rely heavily on synthetic data or purely LLM-as-Judge that fail to capture failure cases. In contrast, GENIUS fills this critical void by being the first benchmark to feature a fully multimodal interleaved context, purely manually curated annotations, and a hybrid evaluation protocol to quantify FI in generative scenarios.

F Details of Method

F.1 Prompt Template for Keyword Generation

To extract task-critical visual cues K_i for each context image I_i, we employ the following prompt to guide Bagel in identifying key regions within the context images. The template is shown in Figure 7. The generated keywords are subsequently used to compute the relevance map.

F.2 Mathematical Formulation of Attention Modulation

In this section, we detail the implementation of the Bias Injection stage. Our strategy is mathematically inspired by Li et al. (2025d;c), while being adapted to our keyword-based relevance scoring mechanism.

The modulation is selectively applied to a subset of decoder layers $\mathcal{L}_{selected}$ and generation steps $\mathcal{T}_{selected}$. For a targeted attention head h in layer $l \in \mathcal{L}_{selected}$ at generation step $t \in \mathcal{T}_{selected}$, let $\mathbf{A}^{l,h} \in \mathbb{R}^{N\times N}$ denote the original attention logits before the Softmax operation. Let $\mathbf{S} \in \mathbb{R}^N$ denote the relevance score vector computed during the Relevance Mapping stage.

Specifically, for each context image I_i, we first extract a set of keywords K_i. The relevance score map S_i is then computed by measuring the similarity between the semantic embeddings of I_i and the token embeddings of K_i. The final relevance map $\mathbf{S}$ is obtained by concatenating all S_i. This process produces a spatial probability map, which is further resized to match the spatial dimensions of the VAE latent embeddings.

To encourage the model to focus on task-critical regions, we inject a dynamic bias term into the attention logits. The modulated attention logits $\hat{\mathbf{A}}^{l,h}$ are computed as:

$$\hat{\mathbf{A}}^{l,h}(i,j) = \mathbf{A}^{l,h}(i,j) + \lambda \cdot \mathcal{F}(\mathbf{S}_j) \tag{9}$$

where i and j denote the query and key token indices, respectively, and λ is a scalar hyperparameter controlling the modulation strength. The transformation function $\mathcal{F}(\cdot)$ normalizes the raw relevance scores into a bipolar bias distribution:

$$\mathcal{F}(\mathbf{S}_j) = \frac{\mathbf{S}_j - \mu_{\mathbf{S}}}{\sigma_{\mathbf{S}} + \epsilon} \tag{10}$$

where $\mu_{\mathbf{S}}$ and $\sigma_{\mathbf{S}}$ denote the mean and standard deviation of the relevance scores within the current context window, and ϵ is a small constant introduced for numerical stability.

The final attention weights are obtained via the standard Softmax operation:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{\hat{\mathbf{A}}}{\sqrt{d}}\right)V \tag{11}$$

This formulation ensures that the gradient norm contribution from noise tokens is effectively dampened by the exponential suppression of the Softmax function.

F.3 Implement Details

For our method, we set the selected generation step range as $\mathcal{T}_{selected} = [10, 35]$ and the selected decoder layer range as $\mathcal{L}_{selected} = [8, 24]$. The modulation strength hyperparameter is set to $\lambda = 1.0$. All experiments are conducted on NVIDIA A100 GPUs.

G THEOREM PART

G.1 EXACT DEFINITION OF $\mathcal{A}$

Let x_t denote the noisy intermediate variable at a certain time step. Let C_1 represent the context and instruction (text modality), and C_2 represent the image modality. Then for $t+1$-th step, we have:

$$x_{t+1} = F(u_1, u_2, g_t) \tag{12}$$

cause Bagel uses MoE architecture, the intermediate variables are defined as:

$$u_1 = \text{Und_Encoder}(C_1 \parallel C_2), \tag{13}$$
$$u_2 = \text{Gen_Encoder}(C_2), \tag{14}$$
$$g_t = \text{Gen_Encoder}(x_t). \tag{15}$$

For the $(l+1)$-th Decoder layer, Bagel employs a **Pre-Layer Normalization (Pre-LN)** structure. Let $u = (u_1||u_2)$ where $\parallel$ denotes matrix concatenation operation. We then obtain the detailed update rule of the decoder layer:

$$g^{(t,l+1)} = \mathcal{A}(u^{(l)}, g^{(t,l)}) + f\left(\text{Up}\left(\mathcal{N}(\mathcal{A}(u^{(l)}, g^{(t,l)})\right)\right) + b = \mathcal{L}^{(l)}_{\text{Up},b}(u^{(l)}, g^{(t,l)}) \tag{16}$$

where the initial bias is set as $b_{\text{initial}} = 0$, RMS denotes the Root Mean Square operation, and Up denotes the Up layer in the decoder block.

The core attention function $\mathcal{A}(u, g)$ is formulated as:

$$\mathcal{A}(u, g) = \text{MoE_attn}\left((U_1 \parallel U_2), G\right) + g \tag{17}$$
$$U_1 = \text{Und}\left(\text{RMSNorm}(u_1)\right) \tag{18}$$
$$U_2 = \text{Gen}\left(\text{RMSNorm}(u_2)\right) \tag{19}$$
$$G = \text{Gen}\left(\text{RMSNorm}(g)\right) \tag{20}$$

where

$$\text{MoE_attn}\left(U, G\right) = \text{Softmax}\left(\frac{G_{\text{query}}\left(U_{\text{key}} \parallel G_{\text{key}}\right)^\top}{\sqrt{d_{\text{attn}}}}\right) \times \left(U_{\text{value}} \parallel G_{\text{value}}\right), \tag{21}$$

where $U = (U_{\text{query}}, U_{\text{key}}, U_{\text{value}})$ and $G = (G_{\text{query}}, G_{\text{key}}, G_{\text{value}})$ denote the query/key/value components of U and G respectively, d_{attn} is the dimension of attention heads. Und (Understanding) and Gen (Generation) denote the Q, K, and V projection layers for different experts in MoE.

G.2 PROOF OF THEOREM 4.1

Our goal is to prove:

$$\mathcal{L}_{\text{Up}+\Delta\text{Up},\, b+\Delta b}(u', g) = \mathcal{L}_{\text{Up},b}(u, g) \tag{22}$$

According to the definition of $\mathcal{L}$, we have:

$$\mathcal{L}_{\text{Up}+\Delta\text{Up},b+\Delta b}(u', g) = \mathcal{A}(u', g) + f((\text{Up} + \Delta\text{Up})(\frac{\mathcal{A}(u', g)}{RMS(\mathcal{A}(u', g))})) + b + \Delta b \tag{23}$$

Cause:

$$\Delta\text{Up} = \frac{\text{Up}\left(\delta\mathcal{A}\right)\mathcal{N}(\mathcal{A}(u', g))^\top}{\left\|\mathcal{N}(\mathcal{A}(u', g))\right\|^2}, \Delta b = \mathcal{A}(u, g) - \mathcal{A}(u', g) \tag{24}$$

We proceed to expand Equation (23):

$$\mathcal{L}_{\text{Up}+\Delta\text{Up},b+\Delta b}(u', g) = \mathcal{A}(u, g) + f(\text{Up}\mathcal{N}(\mathcal{A}(u', g)) + \text{Up}(\delta\mathcal{A})) + b \tag{25}$$
$$= \mathcal{A}(u, g) + f(\text{Up}\mathcal{A}(u, g))) + b \tag{26}$$
$$= \mathcal{L}_{\text{Up},b}(u, g) \tag{27}$$

where

$$\delta\mathcal{A} = \mathcal{N}(\mathcal{A}(u, g)) - \mathcal{N}(\mathcal{A}(u', g)) \tag{28}$$

Thus, we complete the theorem's proof.

G.3 Proof of Theorem 4.2

Our goal is to prove the following conclusion:

$$\begin{cases} \mathrm{Up}_{i+1} = \mathrm{Up}_i - h\nabla_{\mathrm{Up}} L_i(\mathrm{Up}_i), \\ b_{i+1} = b_i - \nabla_b \left(\mathrm{tr}\left(\delta_i^\top b_i\right)\right) \end{cases} \tag{29}$$

For Up, we first expand the update rule directly and obtain:

$$\begin{aligned} \mathrm{Up}_{i+1} - \mathrm{Up}_i &= \Delta\mathrm{Up}_{i+1} - \Delta\mathrm{Up}_i \\ &= \frac{\mathrm{Up}\left(\delta\mathcal{A}_{i+1}\right)\left(\mathcal{N}(\mathcal{A}(g))\right)^\top}{\left\|\mathcal{N}(\mathcal{A}(g))\right\|^2} - \frac{\mathrm{Up}\left(\delta\mathcal{A}_i\right)\left(\mathcal{N}(\mathcal{A}(g))\right)^\top}{\left\|\mathcal{N}(\mathcal{A}(g))\right\|^2} \\ &= \frac{\mathrm{Up}\left(\delta\mathcal{A}_{i+1} - \delta\mathcal{A}_i\right)\left(\mathcal{N}(\mathcal{A}(g))\right)^\top}{\left\|\mathcal{N}(\mathcal{A}(g))\right\|^2} \end{aligned} \tag{30}$$

According to the main text, if we define:

$$\begin{cases} h = \frac{1}{\left\|\mathcal{N}(\mathcal{A}(g))\right\|^2} \\ \Delta_i = \mathrm{Up}\left(\delta\mathcal{A}_i - \delta\mathcal{A}_{i+1}\right)\left(\mathcal{N}(\mathcal{A}(g))\right)^\top \\ L_i(\mathrm{Up}) = \Delta_i^\top \mathrm{Up} \end{cases} \tag{31}$$

and notice the following trivial property:

$$\nabla_{\mathrm{Up}} \,\mathrm{trace}\left(\Delta_i^\top \mathrm{Up}\right) = \Delta_i \tag{32}$$

then we can conclude that:

$$\mathrm{Up}_{i+1} = \mathrm{Up}_i - h\nabla_{\mathrm{Up}} L_i(\mathrm{Up}_i) \tag{33}$$

For b, we have:

$$b_{i+1} - b_i = \Delta b_{i+1} - \Delta b_i = \mathcal{A}(u^{(i+1)}, g) - \mathcal{A}(u^{(i)}, g) \tag{34}$$

so according to property Equation (32):

$$b_{i+1} = b_i - \nabla_b \left(\mathrm{tr}\left(\delta_i^\top b_i\right)\right) \tag{35}$$

Thus, we complete the theorem's proof.

H Limitation and Discussion

While GENIUS provides a robust evaluation paradigm for multimodal generation, we acknowledge two primary limitations. First, our GFI taxonomy is an operationalized framework tailored specifically for visual generation rather than an exhaustive model of human fluid intelligence. Nevertheless, it serves as a minimal functional set; more complex capabilities, such as multi-step compositional planning, can be fundamentally decomposed into these three core dimensions. Second, cross-model quantitative comparisons are influenced by input format discrepancies, as models were evaluated using either interleaved or decoupled inputs based on their inherent compatibility. Rather than a benchmark flaw, this disparity reflects a pervasive architectural limitation—most current Unified Multimodal Models (UMMs) lack robust support for true fine-grained interleaved contexts. While this sensitivity to input formatting inherently underscores the exact deficit in fluid adaptability that GENIUS seeks to expose, we anticipate that this format-induced variance is transitory. As the field increasingly converges toward natively interleaved architectures, this formatting constraint will naturally resolve, enabling fully standardized, format-agnostic evaluations in the future.

I Broader Impacts

This work encourages a fundamental paradigm shift in generative AI, steering the community's focus away from static knowledge retrieval (Crystallized Intelligence) toward dynamic, context-aware reasoning (Fluid Intelligence). By formally operationalizing and quantifying Generative Fluid Intelligence, GENIUS provides a critical stepping stone toward the development of true visual foundation models. Rather than acting as mere probabilistic engines that regurgitate pre-training

distributions, next-generation models equipped with robust fluid intelligence will function as adaptable cognitive systems capable of genuine general-purpose visual reasoning. This evolution promises to catalyze broader innovations across diverse domains. In creative industries, it unlocks deeply personalized, intent-driven synthesis where models can faithfully execute unprecedented visual combinations. In scientific and educational fields, it empowers AI to visualize novel hypotheses, abstract constraints, or counter-intuitive phenomena that intentionally contradict established priors. Ultimately, GENIUS contributes to the broader prosperity of the AI ecosystem by establishing a rigorous standard for intent-aligned generation, ensuring that the future of generative models is built on a foundation of reliable, logical adaptability rather than superficial aesthetic appeal.

Prompt Template for Keyword Generation

You are an **EXPERT IMAGE GENERATION PLANNER**. Your goal is to parse multimodal instructions and map every provided image to its specific role in the generation process.

Task:

Analyze the user's Instruction and the list of image num provided images. Determine precisely what information needs to be extracted or retained from each image.

Focus Definition Rules:

1. **TARGET CANVAS / BASE IMAGE**
   - **Target Canvas / Base Image:** If an image serves as the foundation to be edited, reshaped, or modified (e.g., "change the background of this image", "add a hat to this person"), the value must be "**all**".
   - **Feature Extraction:** If only a specific part is needed (e.g., "swap face", "use this shirt", "holding this object"), output the specific noun (e.g., "face", "shirt", "cup").
   - **Style/Attribute Reference:** If the image provides abstract attributes (e.g., "use this lighting", "copy this art style", "follow this pose"), output the attribute name (e.g., "lighting", "art style", "pose").
   - **Irrelevant:** If an image is mentioned but contributes no visual content to the result, output empty string "".

Output Format:

Return **ONLY** a strictly valid JSON object:
{ "<image X>": "focus_string" }

Few-Shot Examples:

Example 1:
Input: "Transfer the style of <image 1> to the car in <image 2>, but make sure the background matches <image 3>."
Output: {
"<image 1>": "art style",
"<image 2>": "car",
"<image 3>": "background"
}

Example 2:
Input: "Take <image 1> and remove the person from it."
Output: {
"<image 1>": "all"
}

Current Input Context:

- **Images Count:** {image_num}
- **Instruction:**

"""
{content}
"""

Figure 7: **Prompt Template for Keyword Generation.**

Rule Compliance Evaluation

Your task is to act as a **PRECISION VISUAL AUDITOR** in Zero-Tolerance Fidelity Mode.

Mission Statement:

Rigorously evaluate the alignment between the **HINT** and the **MODEL_OUTPUT_IMAGE**. You must prioritize Technical Precision over Perceptual Plasticity. This mode is designed for high-stakes instruction following where "close enough" is considered a failure.

Scoring Hierarchy:

1. **SCORE 2 [PERFECT EXECUTION]**
   - **Standard:** The image is a flawless visual manifestation of the HINT. Every explicit and implicit constraint must be met with 100% accuracy.
   - **Identity and Nouns:** All requested subjects are present, anatomically/structurally correct, and positioned exactly as described.
   - **Adjective/Attribute Fidelity:** Every single descriptor (color, texture, material, style, state, quantity) is rendered without deviation.
   - **Strict Detail Check:** If the hint specifies "three buttons" and there are two or four, it is NOT a 2. If the hint says "crimson" and the output is "bright red," it is NOT a 2.
   - **Rule:** Award only if there is **ZERO** discrepancy between text and pixels.
2. **SCORE 1 [PARTIAL COMPLETION / ACCURACY DRIFT]**
   - **Standard:** The core subject/action is present, but the execution fails on specific details, modifiers, or secondary constraints.
   - **Minor Omissions:**A secondary adjective is ignored (e.g., "vintage" style is missing, but the object is there).
   - **Count/Scale Errors:** Wrong number of objects or incorrect relative sizing.
   - **Color Drift:** The color is in the correct family but lacks the specific shade or intensity requested.
   - **Rule:** Award 1 if the viewer can tell what was intended, but the "fine print" of the instruction was neglected.
3. **SCORE 0 [FAILURE / GROSS ERROR]**
   - **Standard:** Total failure to execute the primary intent.
   - **Subject Error:** The wrong object was added, the target was removed, or the primary subject is unrecognizable.
   - **Non-Action:** No change was made to the image despite a request for modification.
   - **Semantic Inversion/Incoherence:** The model did the opposite of the hint or produced a visual hallucination/glitch.
   - **Rule:** Award 0 if the primary goal of the HINT is unfulfilled.

Evaluation Steps:

1. **SCAN:** List every Noun, Adjective, Count, and Spatial relation in the HINT.
2. **VERIFY PRIMARY:** Is the core action/subject present? (If No → 0).
3. **VERIFY EXHAUSTIVE:** Check every item from Step 1. Any error → 1.
4. **FINAL SCORE:** Award 2 ONLY if Step 3 yields a perfect match.

Evaluation Data:

HINT: {{hint}}
TARGET: {{output_image_tag}}

Output Format:

```
Return ONLY the following line:
Rule Compliance:  X
```

Visual Consistency Evaluation

Your task is to act as a **VISUAL FIDELITY ANALYST** in Strict Evaluation Mode.

Mission Statement:

Your mission is to evaluate the consistency between the REFERENCE IMAGE and the TARGET based on the HINT. Your goal is to distinguish between "Great Work" (2), "Acceptable Effort" (1), and "Total Failure" (0).

Scoring Hierarchy:

1. **SCORE 2 [HIGH FIDELITY (SUCCESSFUL)]**
   - **Standard:** The TARGET is a high-quality implementation of the HINT. The core subject remains stable and the image feels professional.
   - **Strong Identity:** The main subject (person, object, or scene) is clearly the same as in the Reference.
   - **Smooth Transformation:** The changes requested by the HINT are integrated.
   - **Minor Tolerance:** Small shifts in color, slight facial softening, or minor background variations are PERFECTLY ACCEPTABLE.
   - **Rule:** If the image is good, and follows the HINT, give it a 2.
2. **SCORE 1 [RECOGNIZABLE DERIVATION]**
   - **Standard:** The TARGET is clearly related to the Reference, even if the execution is imperfect. This is the catch-all category for images that "get the idea right" but lose some detail.
   - **Recognizable Link:** You can still tell it is based on the same subject or concept, even if the face looks a bit different or background have shifted.
   - **Moderate Drift:** The HINT was attempted, but the model may have simplified original details or introduced AI blurring/messiness.
   - **High Tolerance:** Even if the image has lost some "Visual DNA," as long as it isn't a completely different subject, it stays in this category.
   - **Rule:** If the model tried to follow the HINT and the result is "okay" or "recognizable," award a 1.
3. **SCORE 0 [TOTAL FAILURE]**
   - **Standard:** Total failure to execute the primary intent or maintain subject connection.
   - **Subject Swap:** The model generated a completely different person or object.
   - **Ignored Instruction:** The model provided a generic image that ignores both the Reference and the HINT entirely.
   - **Broken Output:** The image is a corrupted, unidentifiable mess of pixels.
   - **Rule:** Do NOT award a 0 if there is any link to the Reference Image.

Evaluation Steps:

1. **RELATION:** Is the Target even remotely related to the Reference subject? (If No → Score 0).
2. **QUALITY & FIDELITY:** Does the Target look stable, professional, and closely follow the HINT without distracting errors? (If Yes → Score 2).
3. **DRIFT ASSESSMENT:** For everything else in between (minor drift, detail loss, background shifts, but same subject) → Score 1.

Evaluation Data:

> **REFERENCE IMAGE:** {{reference_image}}
> **HINT:** {{hint}}
> **TARGET:** {{output_image_tag}}

Output Format:

```
Return ONLY the following line:
Visual Consistency:  X
```

Aesthetic Quality Evaluation

Your task is to act as a **PROFESSIONAL VISUAL ARTS CURATOR** in High-Visual Harmony Mode.

Mission Statement:

Your mission is to categorize images into Masterpiece Level (2), Standard Work (1), and Technical Failure (0). You must reward images that achieve high visual harmony and logical consistency, prioritizing Visual Cohesion and professional merit.

Scoring Hierarchy:

1. **SCORE 2 [EXCEPTIONAL / MASTERPIECE LEVEL]**
   - **Standard:** High-end, commercial quality. The image is indistinguishable from professional work.
   - **Structural Logic:** All objects and characters follow the laws of physics and anatomy.
   - **Rendering Clarity:** Textures look intentional and sharp; lighting creates a convincing sense of depth.
   - **Visual Appeal:** The overall composition is professional and free of distracting AI hallucinations.
   - **Tolerance Clause:** Award a 2 even if there is a tiny, non-distracting flaw, provided the overall impact is professional.
2. **SCORE 1 [STANDARD / THE GOOD EFFORT ZONE]**
   - **Standard:** The DEFAULT category. Visually acceptable and logically sound but contains visible "AI-isms."
   - **Visible Flaws:** Slightly soft hands, plastic skin, or shadows that do not perfectly align with the light source.
   - **Minor Perspective Issues:** Background elements that are slightly tilted or out of scale.
   - **Rule:** Use this as the safety score for any image that is "pretty good" but not perfect.
3. **SCORE 0 [FAILED / TECHNICAL FAILURE]**
   - **Standard:** Images that are visually nonsensical or structurally broken.
   - **Logical Collapse:** Extra limbs that are gross and distracting, or faces that have lost basic human structure.
   - **Extreme Noise:** Grain or artifacts so heavy they obscure the main subject.
   - **Rule:** Only award 0 if the image is unusable. If the parts are mostly in the right place, do NOT give a 0.

Evaluation Steps:

1. **SCAN:** Identify the visual logic, realism, and presence of any AI-generated artifacts.
2. **VERIFY LOGIC:** Does the image maintain basic structural integrity? (If No → 0).
3. **ASSESS QUALITY:** Check for professional rendering, sharp textures, and lighting depth.
4. **FINAL SCORE:** Award 2 for commercial quality, 1 for standard AI output, and 0 for unusable failure.

Evaluation Data:

> **METRIC:** Aesthetic Quality (Visual Logic and Realism)
> **TARGET:** {{output_image_tag}}

Output Format:

```
Return ONLY the following line:
Aesthetic Quality:  X
```

Figure 10: **Prompt Template for Aesthetic Quality Evaluation.**



Figure 11: **Detailed Qualitative Examples and Model Outputs. (1/2)**



Figure 12: **Detailed Qualitative Examples and Model Outputs. (2/2)**