Unify-Agent: A Unified Multimodal Agent for World-Grounded Image Synthesis

Shuang Chen[1,2], Quanxin Shou[4], Hangting Chen[2], Yucheng Zhou[2], Kaituo Feng[3]
Wenbo Hu[1], Yi-Fan Zhang[3], Yunlong Lin[3], Wenxuan Huang[3], Mingyang Song[3]
Dasen Dai[3], Bolin Jiang[4], Manyuan Zhang[3], Shi-Xue Zhang[2], Zhengkai Jiang[2]
Lucas Wang[2], Zhao Zhong[2], Yu Cheng[3], Nanyun Peng[1]

[1]University of California, Los Angeles [2]Tencent Hunyuan
[3]The Chinese University of Hong Kong
[4]The Hong Kong University of Science and Technology

Abstract

Unified multimodal models provide a natural and promising architecture for understanding diverse and complex real-world knowledge while generating high-quality images. However, they still rely primarily on frozen parametric knowledge, which makes them struggle with real-world image generation involving long-tail and knowledge-intensive concepts. Inspired by the broad success of agents on real-world tasks, we address this limitation through agentic modeling. Specifically, we present **Unify-Agent**, a unified multimodal agent for *world-grounded image synthesis* that reframes image generation as a four-step agentic pipeline—prompt understanding, multimodal evidence searching, grounded recaptioning, and final synthesis—all implemented within a single unified multimodal backbone. To train our model, we construct a tailored multimodal data pipeline and curate 143K high-quality agent trajectories for world-grounded image synthesis, enabling effective supervision over the full agentic generation process. We further introduce FactIP, a benchmark covering 12 categories of culturally significant and long-tail factual concepts that explicitly requires external knowledge grounding. Extensive experiments show that our proposed Unify-Agent substantially improves over its base unified model across diverse benchmarks and real world generation tasks, while approaching the world knowledge capabilities of the strongest closed-source models. As an early step toward agent-based world-grounded image synthesis, our work highlights the value of tightly coupling reasoning, searching, and generation within a unified backbone.

1 Introduction

Recent advances in Text-to-Image (T2I) generation have significantly improved visual quality and controllability [36, 37, 35, 43, 24]. Yet a growing challenge is whether such models can faithfully depict entities grounded in the real world, including real people, cultural symbols, rare intellectual properties (IPs), historical scenes, scientific phenomena, and other long-tail concepts, where success demands factual fidelity rather than mere visual plausibility [8, 38]. Unified multimodal models (UMMs) offer a natural approach by unifying understanding and generation within a shared architecture [44, 55, 54, 9, 13, 62, 50, 7], but their world knowledge is locked in fixed parameters and cannot be updated at inference time. As a result, the bottleneck for long-tail prompts is thus not rendering quality but *knowing what the target should look like*, suggesting that factual synthesis benefits from external evidence at inference rather than parametric memory alone.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

A natural remedy is to allow the model to access external world knowledge during inference instead of relying solely on parametric memory. Existing agentic text-to-image systems pursue this direction, but most are multi-stage pipelines that loosely connect an LLM planner, retrieval tools, and a standalone image generator [51, 5, 42, 57], leaving evidence acquisition, multimodal reasoning, and visual synthesis as separate stages rather than parts of a unified generative process. Moreover, retrieved text rarely specifies fine-grained visual attributes, while reference images often introduce irrelevant content that conflicts with user-specified composition [3, 31]. The key challenge is therefore not simply retrieving external knowledge, but turning it into visual guidance that preserves fidelity to the source while remaining aligned with user intent.

To this end, we present Unify-Agent, a unified multimodal agent for world-grounded image synthesis that implements four capabilities within a single backbone: (1) *THINK*, for prompt understanding and information-gap detection; (2) *RESEARCH*, for actively acquiring relevant textual and visual knowledge from external sources; (3) *RECAPTION*, for converting retrieved knowledge into structured textual guidance; and (4) *GENERATE*, for final image synthesis. Rather than passing raw retrieved content to the generator, Unify-Agent uses recaptioning as a intelligent filtering mechanism that transforms external knowledge into structured constraints, disentangling identity-preserving cues from scene-compositional requirements. External tools are used only during training-data construction and as standard search APIs at inference; the agentic policy itself is realized within a single backbone trained end-to-end via supervised trajectory distillation.

To realize this vision, we build the agentic workflow on a unified model rather than a decoupled pipeline, allowing the same backbone to handle both retrieval-grounded reasoning and image synthesis. A practical benefit of this design appears in the reception stage: because the unified backbone exposes both high-level ViT [14] semantic tokens and low-level VAE [23] latents of retrieved reference images, the language branch has access to richer visual evidence than a text-only LLM. Our ablations support this in a narrow form: removing the ViT pathway hurts reception quality the most, while VAE features provide a smaller but consistent additional benefit. We treat this as evidence that joint access to semantic and appearance features helps recaptioning. To evaluate this capability, we introduce FactIP, a benchmark covering 12 categories of rare identities and culturally significant long-tail concepts. Experimental results show that Unify-Agent substantially improves factual visual synthesis: on FactIP, it attains an overall score of 73.2, surpassing its base model (Bagel [12]) by more than 22 points and outperforming generation-only baselines such as FLUX.1-dev [24] and SD-3.5-large [15]. Unify-Agent also achieves leading performance among open-source unified models on world-knowledge benchmarks including WiSE [33], KiTTEN [18], and T2I-FactualBench [22].

In summary, our main contributions are as follows:

- ◇ **A Unified Agentic Pipeline for World-Grounded Synthesis.** We introduce Unify-Agent, which reformulates text-to-image generation as an inference-time sequential process (Think, Research, Reception, Generate) realized within a single unified multimodal backbone.
- ◇ **Empirical Analysis of Reception Inside a Unified Backbone.** We analyze how the joint availability of high-level semantic (ViT [14]) and low-level appearance (VAE [23]) features inside a unified model affects evidence recaptioning, and find that ViT features dominate while VAE features provide a smaller, complementary benefit.
- ◇ **A Comprehensive Benchmark.** We present FactIP, a curated benchmark specifically designed to evaluate identity consistency and factual faithfulness of world-grounded generation for rare and long-tail concepts.
- ◇ **Strong Empirical Performance.** Unify-Agent achieves leading results among open-source unified models on several factual benchmarks [33, 18, 22] (FactIP, WiSE, KiTTEN, and T2I-FactualBench) and narrows the gap to leading commercial models on world-knowledge generation.

2 Related Works

Unified multimodal models. Recent unified multimodal models (UMMs) [44, 62, 13, 47, 20] aim to jointly support visual understanding and image generation within a shared backbone. Architectures like the Janus family [54, 9] and Show-o [55] have demonstrated that semantic comprehension and continuous visual synthesis can be effectively harmonized. Despite this, current UMMs remain strictly *closed-book* systems. Because they rely exclusively on static parametric memory internalized during pre-training, they frequently hallucinate or suffer from identity drift when prompted with

rare, long-tail, or world-dependent entities [22, 6, 10]. This bottleneck highlights the need to shift from purely parametric generation toward an *open-book* approach, where the model actively acquires external evidence before synthesis.

Agentic workflows. Agentic reasoning has substantially advanced large language models through iterative planning and tool use [59, 61, 21]. Recent efforts have extended similar pipelines to text-to-image generation [51, 5, 42], but most adopt a *multi-API stitching* paradigm [41, 53, 56, 58, 17] that chains frozen text-only planners, search tools, and standalone image generators. This decoupled design tends to accumulate cascading errors and keeps multimodal reasoning separate from visual synthesis. Unify-Agent instead realizes the agentic loop—gap detection, cross-modal evidence grounding, and generation—within a single unified backbone, reducing reliance on brittle API pipelines.

World knowledge and factual evaluation. Conventional text-to-image benchmarks predominantly emphasize aesthetic quality and generic prompt alignment [19, 26], largely overlooking the factual correctness of generated content. To address this, recent efforts have pivoted towards knowledge-oriented evaluation. Notably, T2I-FactualBench [22] provides a comprehensive diagnosis of generative models, systematically exposing their struggles with knowledge-intensive concepts across multiple domains. This severe deficiency in explicit world knowledge is further echoed by related benchmarks such as KITTEN [18] and WISE [33]. Nevertheless, existing evaluations remain fundamentally *diagnostic*—they quantify the knowledge gap but offer no structural remedy. Motivated by these insights, we not only construct a targeted evaluation for rare identities and long-tail concepts, but also propose an evidence-grounded agentic framework that actively bridges this gap during synthesis.

3 Preliminary

3.1 Unified Multimodal Model: Bagel

We build our Unify-Agent upon Bagel [12], which employs a Mixture-of-Transformers (MoT) [28] architecture with shared self-attention to unify visual understanding and continuous image generation in a single framework. Bagel disentangles these two capabilities through dedicated experts:

Multimodal Understanding. Handled by a dedicated understanding expert, the understanding pathway is formulated as autoregressive next-token prediction via a language modeling head. Conditioned on the multimodal input context C, the training objective minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{text}} = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}, C)\,, \tag{1}$$

where x_t is the target text token, $x_{<t}$ denotes the preceding token sequence, and C encapsulates the provided multimodal context.

Multimodal Generation. Conversely, handled by a generation expert, the generation pathway is formulated as a rectified flow [30] operating within the latent space of a continuous VAE [25]. Conditioned on the same multimodal context C, the model learns a time-conditioned velocity field u_θ by minimizing the latent flow-matching objective:

$$\mathcal{L}_{\text{image}} = \mathbb{E}_{t\sim\mathcal{U}(0,1),z_t}\,\|u_\theta(z_t, t\,;\,C) - u^\star(z_t, t)\|_2^2\,, \tag{2}$$

where $t \sim \mathcal{U}(0,1)$ is the continuous timestep, z_t represents the latent state at time t, $u^\star(z_t,t)$ is the target vector field, and $u_\theta(z_t, t; C)$ is the velocity field predicted by the generation expert.

3.2 Motivating Evidence for World-Grounded Synthesis

The failures of world-grounded image synthesis on long-tail IPs stem from missing world knowledge rather than insufficient visual fidelity: base T2I models often do not know a rare concept's appearance, identity-defining attributes, or how to instantiate them under compositional instructions. This motivates inference-time knowledge injection, where text supplies semantics and attributes while images provide appearance anchors.

To validate this, we conduct a preliminary training-free study on 200 FactIP examples (*scene*, *character*, *object*), comparing Bagel under four settings: **prompt-only**, **text injection** (appending IP background), **visual injection** (two ground-truth reference images), and **text + visual injection**. As shown in Figure 1, both modalities improve over the baseline, with visual injection yielding substantially larger gains, supporting proactive retrieval at inference.

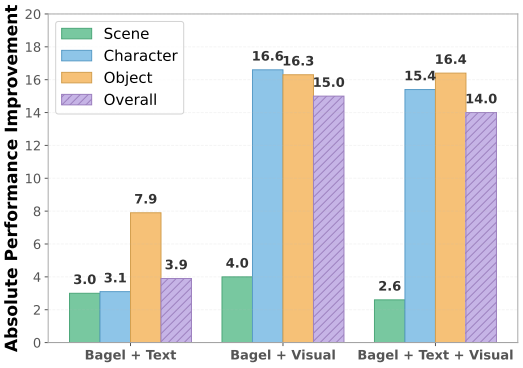


Figure 1: Absolute performance gains over prompt-only **Bagel** on rare-IP factual generation under different knowledge injection strategies, across three FactIP categories (Scene, Character, Object). Purple hatched bars denote the overall average gain.

Figure 1 also reveals the limitation of naive knowledge injection. Though combining text and visual inputs remains beneficial, its improvement is slightly weaker than visual injection alone, suggesting that simply concatenating retrieved text with prompts does not optimally exploit multimodal evidence. Raw text injection may introduce redundant or weakly visual information that interferes with instruction following, while raw visual injection, though effective at anchoring identity, may over-constrain generation and reduce compositional flexibility. These findings motivate our reception paradigm, which distills retrieved textual and visual knowledge into a unified, structured, and generation-oriented description for world-grounded synthesis.

3.3 Problem Formulation

Standard T2I models $p_\theta(y \mid x)$, synthesizing an image y from a prompt x using only the parametric memory θ. For rare entities, this is intractable due to knowledge gaps in θ; yet as shown in Section 3.2, naively expanding the condition to $p_\theta(y \mid x, \mathcal{K}_{\text{text}}, \mathcal{K}_{\text{vis}})$ with raw retrieved knowledge $\mathcal{K}$ is also suboptimal.

We thus formulate world-grounded synthesis as an **interleaved generative trajectory**, introducing four intermediate variables—the cognitive gap assessment g, textual evidence trace τ_t, visual evidence trace τ_v, and evidence-grounded reception c—that yield the joint distribution:

$$p_\theta(y, c, \tau_t, \tau_v, g \mid x) = \underbrace{p_\theta(g \mid x)}_{\text{Gap Detection}} \cdot \underbrace{p_\theta(\tau_t, \tau_v \mid x, g)}_{\text{Evidence Acquisition}} \cdot \underbrace{p_\theta(c \mid x, g, \tau_t, \tau_v)}_{\text{Recaptioning}} \cdot \underbrace{p_\theta(y \mid c, \tau_v)}_{\text{Visual Synthesis}} \tag{3}$$

Unifying discrete modeling (for g, τ_t, τ_v, c) and continuous latent modeling (for y) under a shared θ, this factorization defines the four cognitive phases of **Unify-Agent**:

1. **Cognitive Gap Detection** ($p_\theta(g \mid x)$): The model evaluates x against θ to identify missing visually-critical attributes, yielding a reasoning trace g.
2. **Evidence Acquisition** ($p_\theta(\tau_t, \tau_v \mid x, g)$): Conditioned on g, the agent retrieves and reasons over textual evidence τ_t, then visual evidence τ_v.
3. **Evidence-Grounded Recaptioning** ($p_\theta(c \mid x, g, \tau_t, \tau_v)$): The model consolidates the evidence and the original instruction into a structured, generation-oriented reception c.
4. **Visual Synthesis** ($p_\theta(y \mid c, \tau_v)$): The image is generated from only the reception c and visual identity anchors τ_v, approximated via attention masking (Appendix C) that restricts flow-matching tokens to attend to c and τ_v rather than upstream x, g, τ_t. We treat Eq. 3 as a modeling abstraction rather than a strictly enforced independence assumption.

4 Data Pipeline

Training a unified agent for world-grounded synthesis requires data that captures *how* the agent reaches its output, not just the output itself; evaluating it requires a benchmark that genuinely stresses world knowledge rather than memorization. As shown in Figure 2, our pipeline addresses both: (1) trajectory-level training data, and (2) the FactIP evaluation benchmark.

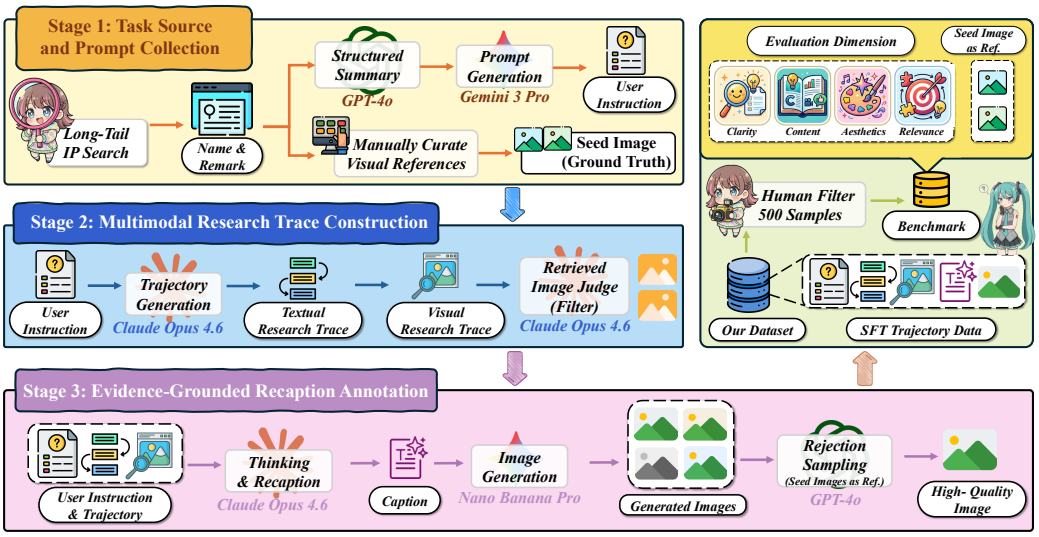


Figure 2: **Overview of our data pipeline.** Starting from long-tail IP collection, we construct user instructions and ground-truth images, build multimodal research trajectories with textual and visual evidence, and perform evidence-grounded recaption annotation. The resulting dataset supports both SFT trajectory learning and the FactIP benchmark, which evaluates generation along clarity, content, aesthetics, and relevance.

4.1 Training Data Construction

Each training sample is a quadruple $\mathcal{D}_{\text{SFT}} = \{(x, \tau_t, \tau_v, c)\}$, pairing the prompt x with the textual research trace τ_t, visual research trace τ_v, and final evidence-grounded recaption c. The key departure from standard instruction tuning is that supervision is *process-aware*: by exposing the full retrieval trajectory rather than only the recaption target, the model learns not just *what* evidence to use, but also *when* and *how* to acquire it.

Task source and prompts. We focus on long-tail IPs because that is where world-knowledge gaps actually surface; common entities are too easy to memorize. Annotators curate IPs from the web across 12 domains (*Celebrity*, *Animation*, *Game*, *Comic*, *Mythology*, *Mascot*, *Animal*, *Food*, *Art*, *Toy*, *Landmark*, *Festival*); for each, BabelNet [32] and Wikipedia [1] provide background, two seed images serve as ground truth, and *GPT-4o* [46] extracts identity-defining metadata. After manual filtering, this yields a pool of **456K** concepts. From this pool, *Gemini 3 Pro* [45] generates prompts that deliberately combine factual identity with creative variation in pose, costume, scene, or lighting—ensuring the task tests both grounding and compositional flexibility, not retrieval-style copying.

Multimodal research traces. Trace construction is sequential by design. Textual evidence is acquired to disambiguate the entity, then used to anchor a more targeted visual search, since a visual query issued from the prompt alone has no semantic anchor and tends to surface look-alikes. We use *Claude Opus 4.6* as teacher agent for both stages: what we want the model to learn is not a fixed retrieval recipe but *when* to stop searching and *how* to refine queries, behavior that only emerges when the supervisor itself reasons over partial evidence. Given x and inferred cognitive gap g, the agent issues a textual query q_t to obtain E_t (identity disambiguation, categorical background, role-specific context), then a visual query q_v conditioned on (x, g, τ_t) to retrieve a candidate set $\widetilde{E}_v$. Image-search results, however, routinely contain noises such as watermarks, off-target subjects, or cluttered backgrounds, so candidates are scored by *Gemini 3 Flash* [45] on four axes (identity consistency, subject salience, image clarity, watermark cleanliness), and only the top-2 are retained as E_v. The final trace $(\tau_t, \tau_v) = (q_t, E_t, q_v, E_v)$ aligns directly with the factorized term $p_\theta(\tau_t, \tau_v \mid x, g)$ in Eq. 3, supervising both *what to retrieve* and *how to query for it*.

Evidence-grounded recaption with generation-based validation. A trace that looks plausible to a reader is not necessarily useful for a generator—the recaption is what bridges the two. Following

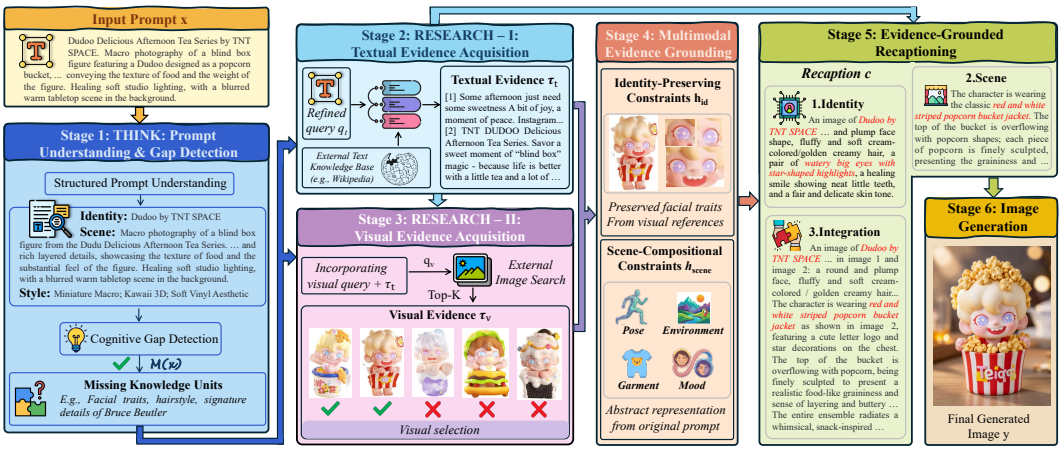


Figure 3: **Overview of the agentic pipeline.** Given a prompt, the model first performs cognitive gap detection, then sequentially acquires textual and visual evidence. Retrieved evidence is decomposed into two grounded constraints—identity-preserving and scene-compositional—which an evidence-grounded recaption fuses into the final specification for image synthesis.

recent prompt-rewriting work [2, 49, 4], we synthesize $c = \mathcal{C}(x, E_t, E_v)$ that fuses three signals: prompt-level composition, textual context, and identity cues from references. To verify each c is operationally useful but not merely semantically correct, we close the loop: *Nano Banana Pro* [45] renders $\hat{y}$ from (c, E_v), and *GPT-4o* [46] judges identity faithfulness against the ground-truth IP image. Trajectories failing five rounds of rejection sampling [29] are discarded as unreliable. This filter is what turns the raw trace pool into **143K** high-quality trajectory–image pairs.

4.2 FactIP Evaluation Benchmark

FactIP is built from a held-out subset of the same long-tail pool, deliberately disjoint from training data. Beyond standard difficulty, we filter for samples that satisfy four properties simultaneously: knowledge-intensive (so success demands real world knowledge), visually grounded (so identity is checkable against references), hard for memorization-only models, and diverse across categories. Manual filtering yields 2,500 samples; from these, we sample a 500-instance test split preserving the category distribution. Each generation is scored along **Clarity**, **Content**, **Aesthetics**, and **Relevance**, with *Relevance*—identity preservation under the instruction and reference—carrying the most weight, since it directly measures the world-grounding capability the benchmark is designed to test. Full construction and scoring details are in Appendix E.1 and Appendix D. We further validate MLLM-based factuality evaluation with a human preference study on FactIP-Mini, focusing on identity preservation and prompt faithfulness (Appendix E.3).

5 Methodology

5.1 Unified Fine-Tuning on Multimodal Agent Trajectories

We adapt Bagel (Section 3.1) to the agentic setting via supervised fine-tuning on the interleaved trajectories from Section 4.1. The challenge is that each trajectory mixes textual reasoning, tool calls, recaptions, and image-generation targets, requiring a single backbone to optimize discrete and continuous objectives jointly. We address this with a dual-loss objective $\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{image}}$, sequence packing across heterogeneous samples, and a hybrid attention mask that prevents upstream reasoning noise from leaking into the synthesis stage.

For text tokens (reasoning, tool invocation, recaption), we use the autoregressive cross-entropy loss from Section 3.1, with one targeted modification: special boundary tokens (`<think>`, `<tool_call>`, `<recaption>`) are upweighted, since reliable structural formatting is what makes the trajectory executable at inference. For image tokens, we retain Bagel's latent flow-matching objective. The two losses are activated by separate supervision masks: text-only segments contribute

only to $\mathcal{L}_{\text{text}}$, while image-generation segments contribute to both. Detailed masking rules are deferred to Appendix C.

5.2 Unify-Agent: Inference-Time Multimodal Agent Pipeline

Unify-Agent instantiates the factorization in Eq. 3 as a sequential decision process (Figure 3): given x, the model decides whether external knowledge is needed, acquires it, distills it into generation-ready constraints, and only then synthesizes the image.

THINK: Prompt understanding and cognitive gap detection. The agent treats x not as a self-sufficient description but as a *partially specified* request. It decomposes the prompt into structured factors (identity, scene, style, photographic intent) and assesses whether visually critical attributes are reliably present in its parametric memory, producing a free-form reasoning trace $g \sim p_\theta(g \mid x)$ that names the missing knowledge units (e.g., facial structure, signature appearance, scene-specific facts). If g is non-trivial, the agent enters the research stage, otherwise it proceeds directly to generation.

RESEARCH: Sequential multimodal evidence acquisition. Conditioned on g, the agent acquires evidence according to $p_\theta(\tau_t, \tau_v \mid x, g)$ in two stages, *textual before visual*, since a visual query without semantic grounding tends to surface look-alikes or wrong-context images of the same name. The agent first issues a textual query to obtain τ_t (identity disambiguation, background, role-specific context), then formulates a visual query conditioned on (x, g, τ_t) to retrieve τ_v, a small set of identity-relevant, context-compatible references.

RECAPTION: Multimodal grounding into executable specifications. The central design choice of Unify-Agent is that retrieved evidence is *not* passed to the generator in raw form. As Section 3.2 shows, raw text injects weakly visual content that hurts instruction following, while raw images over-constrain composition. The recaption stage instead reorganizes the evidence into two disentangled constraints: an *identity-preserving* constraint that fixes visual attributes tied to the target (facial structure, appearance cues, object-defining features), and a *scene-compositional* constraint that encodes prompt-specified pose, environment, garment, and mood. The agent fuses both into the final specification:

$$c \sim p_\theta(c \mid x, g, \tau_t, \tau_v). \tag{4}$$

This separation is what lets recaption preserve identity fidelity *and* compositional controllability simultaneously—the failure mode that defeats naive injection.

GENERATE: Evidence-grounded image synthesis. The generator finally synthesizes

$$y \sim p_\theta(y \mid c, \tau_v), \tag{5}$$

implemented via the hybrid attention mask in Appendix C, which restricts flow-matching tokens to attend only to c and τ_v rather than the full upstream history (x, g, τ_t). We treat this as an inductive bias rather than a strict independence assumption: it shields the continuous generative process from noisy reasoning traces and redundant text, while preserving the complementary roles of c (controllable description) and τ_v (identity anchor).

6 Experiments

6.1 Experimental Details

Baselines We compare Unify-Agent against three categories of baselines. (i) **Commercial models**, including Seedream-Family[40], Nano Banana-2[16], DALLE-3 [2], and GPT-Image-1.5[34], serving as upper-bound references. (ii) **Generation-only models**, including FLUX.1-dev [24], SD-3.5-large [15], Playground-v2.5[27], Z-Image[48], and Qwen-Image [52]. (iii) **Unified MLLMs**, including Janus-Pro-7B [9], Emu3.5[11], Echo-4o[60], Hunyuan-Image-3.0[4], and Bagel [13], which are structurally comparable open-source models that natively unify understanding and generation. Together, this suite covers both proprietary and open-source models across major paradigms.



Figure 4: Qualitative comparison of multi-image generation results on knowledge-intensive prompts involving historical figures, fictional characters, products, and stylized toys. Our method consistently produces images that better preserve subject identity, fine-grained attributes, and prompt-specific details, while achieving stronger real-world knowledge grounding than competing baselines, including Flux-1, Bagel-7b, Hunyuan, and Stable Diffusion.

Evaluation Setup We adopt the *MLLM-as-a-Judge* paradigm [6], which has emerged as a scalable proxy for expert annotation in fine-grained multimodal evaluation. For WISE [33], T2I-FactualBench [22], and KiTTEN [18], we follow prior protocols and use ChatGPT-4o [46] as the evaluator to ensure consistency. For our FactIP benchmark, we use Seed2.0 [39] given its stronger multimodal judging capability. Full evaluation prompts and scoring criteria are provided in Appendix D.

6.2 Main Results

Main Results. Across all four benchmarks, Unify-Agent consistently leads the unified-MLLM group and narrows the gap to commercial models. On **FactIP** (Tab. 1), it attains the best Overall score of 73.2, driven by top Relevance on Character (67.3), Object (71.8), and Scene (78.2)—far above generation-only baselines such as FLUX.1-dev (28.9) and SD-3.5-large (27.5), indicating that explicit multimodal reasoning and external knowledge are essential beyond scaling diffusion priors. On **WiSE** (Tab. 4), it reaches an Overall WiScore of 0.77, surpassing BAGEL+CoT (0.70) with the strongest cultural (0.82), biological (0.72), and chemistry (0.70) results. On **KiTTEN** (Tab. 5), it further sets a new SOTA of 4.08 (text 4.22, entity 3.93). On **T2I-FactBench** (Tab. 6), it leads SKCI (77.4) and MKCC (71.5), with SKCM (69.2) competitive even with DALLE-3 (55.5). Full per-category tables for all three benchmarks are deferred to Appendix A. Qualitative results in Fig. 4 likewise show stronger identity preservation and fewer missing attributes than Flux-1, Bagel-7B, Hunyuan, and Stable Diffusion. Together these results show that Unify-Agent reliably preserves

Table 1: FactIP benchmark results with category-level weighted averages. **Character** includes celebrity, animation, game, comic, mythology, and mascot prompts; **Object** includes animals/plants, food, cultural relic/art, and toy prompts; **Scene** includes landmark and celebration prompts. denotes the best score and the second-best within each group.

Model	Clarity			Content			Aesthetics			Relevance			Overall
	Character	Object	Scene	Character	Object	Scene	Character	Object	Scene	Character	Object	Scene	
Commercial Models													
GPT-Image-1	94.6	94.1	**96.2**	83.5	71.1	78.4	89.2	88.1	94.1	61.9	60.2	76.4	69.3
GPT-Image-1.5	94.4	**94.2**	95.6	82.7	73.2	77.9	89.5	88.1	93.8	62.6	61.5	77.2	69.9
Seedream-4	93.9	83.7	88.0	81.0	92.3	78.1	88.2	81.7	91.5	80.5	90.5	86.0	83.0
Seedream-4.5	94.1	83.1	87.8	80.1	92.2	76.9	88.4	79.7	90.8	81.0	91.3	83.7	82.0
Seedream-5	95.0	86.0	88.1	**86.7**	94.3	83.5	90.3	87.1	91.7	84.3	91.8	87.1	87.3
Qwen-Image-2.0-Pro	94.2	92.7	90.8	77.0	73.6	76.2	86.2	87.5	90.0	62.6	74.0	79.7	71.1
Nano Banana	94.8	71.8	76.7	44.3	**97.8**	84.0	76.0	64.8	93.7	74.7	73.7	52.4	56.6
Nano Banana-Pro	94.3	76.9	86.8	59.1	92.6	69.1	86.7	59.9	91.9	76.9	90.3	72.6	66.7
Nano Banana-2	**96.6**	88.2	90.4	86.3	96.0	**86.6**	**92.2**	**92.2**	**95.3**	**85.5**	**93.9**	**89.5**	**88.5**
Generation Only													
Pixel-Art-XL	34.6	16.9	15.6	14.7	5.7	9.2	28.2	14.7	19.0	10.8	3.7	9.2	12.1
FLUX.1-schnell	85.5	80.6	82.1	35.1	19.9	27.6	68.4	65.5	69.7	13.9	7.9	16.9	24.0
FLUX.1-dev	92.5	89.0	91.8	51.9	40.4	37.7	78.4	75.8	80.3	17.0	8.2	18.2	28.9
FLUX.2-dev	92.4	91.7	90.6	**72.9**	58.6	66.8	84.4	**83.8**	88.5	**46.6**	**46.3**	61.0	**56.3**
SD-3-medium	85.1	81.7	78.5	40.9	22.0	28.6	70.6	64.9	66.8	16.2	6.6	17.9	25.7
SDXL-base-0.9	82.9	78.3	80.2	40.4	16.6	25.3	74.1	67.0	73.2	17.9	6.4	17.1	26.5
SD-3.5-large	85.4	82.5	81.1	44.3	23.2	32.1	75.3	68.1	74.5	18.0	5.9	20.2	27.5
SD-3.5-medium	84.3	73.6	79.2	40.9	17.6	31.6	72.5	54.6	67.4	17.7	7.1	20.2	26.5
Z-Image	94.3	**92.4**	**91.9**	72.5	**62.1**	67.7	**84.9**	83.7	87.3	42.8	44.5	60.5	54.2
Qwen-Image	**94.5**	91.7	91.7	**72.9**	57.3	**67.9**	**84.9**	**83.8**	**89.4**	45.3	43.1	**62.6**	55.4
Unified MLLM													
Janus-1.3B	42.7	24.9	29.7	26.1	13.3	19.2	41.8	26.5	37.1	15.2	10.3	21.0	19.3
Janus-Pro	60.0	40.3	43.9	40.1	25.6	28.5	58.2	46.4	56.6	24.3	21.4	30.2	30.3
Emu3	82.7	81.2	78.6	48.4	29.6	34.8	73.4	69.8	73.7	19.2	13.8	26.1	30.0
Echo4o	89.9	90.7	89.7	65.1	49.3	61.0	80.8	80.5	85.0	36.1	35.7	52.0	47.3
Hunyuan-Image-3.0	93.0	91.1	90.0	72.3	59.7	70.4	84.6	83.0	87.7	40.6	46.8	63.3	53.4
Emu3.5	**94.8**	**92.7**	**94.3**	64.0	54.5	69.3	**85.2**	84.5	**90.6**	49.7	44.4	60.8	57.2
Bagel	91.4	91.8	90.6	68.5	59.1	65.0	81.6	83.3	87.2	39.9	44.0	50.7	50.9
Bagel-CoT	90.7	90.8	90.5	66.4	53.7	62.3	80.3	79.3	85.0	36.2	34.7	47.4	47.0
Unify-Agent (Ours)	92.4	91.5	90.7	**75.8**	**76.1**	**73.6**	83.3	**86.0**	86.4	**67.3**	**71.8**	**78.2**	**73.2**

entity-level details, handles multi-concept compositions, and benefits most in regimes that demand factual recall rather than purely visual priors.

6.3 Analysis of Experimental Results

6.3.1 Ablation Study

Tab. 2 presents ablations on FactIP. The full Unify-Agent improves the Overall score from 50.9 (vanilla Bagel) to 73.2, with the largest gain on **Relevance** (44.9 $\rightarrow$ 72.4)—indicating that the model's main advantage lies in factual grounding and identity fidelity, not generic image quality.

Pipeline components. Removing *text search* drops Overall to 65.4, showing that textual retrieval provides semantic scaffolding for disambiguation. Removing *image search* causes a larger drop (Overall 56.2; Relevance 72.4 $\rightarrow$ 50.8), confirming that visual evidence is critical for fine-grained identity preservation. Removing *recaptioning* also degrades performance substantially (Overall 62.9), indicating that raw retrieved evidence must be reorganized into a compact, executable specification before generation.

Constraint design. Without *identity-preserving* constraints, Relevance falls to 65.9, confirming their role in subject-specific fidelity. Without *scene-compositional* constraints, Content (70.8) and Aesthetics (80.7) decline, showing they are essential for instantiating prompt-specified scene attributes. Overall, the gains of Unify-Agent arise from the joint effect of multimodal evidence acquisition, grounded recaptioning, and structured constraints—no single component is sufficient on its own.

6.4 Discussion: How Unified Visual Representations Support Recaptioning

The reception-architecture ablations in Tab. 2 reveal a clear asymmetry: within a Bagel-style unified backbone, *high-level ViT semantics dominate reception quality, while low-level VAE latents play a smaller complementary role*. The reason is that reception is not a pure language task—it requires identifying *what* appears in retrieved references, *which* attributes are identity-critical, and *how* to express them for downstream synthesis. ViT tokens directly support this semantic reasoning, whereas

Table 2: Ablation study of Unify-Agent on the FactIP benchmark. We report top-level weighted averages over **Clarity**, **Content**, **Aesthetics**, and **Relevance**. The middle rows show expected performance drops when removing key pipeline components, grounded constraint types, or recaption-stage architectural modules. Numbers in parentheses denote changes relative to the vanilla Bagel baseline; red indicates improvement and green indicates degradation. The best result in each column is marked in **bold**, and the second-best is underlined.

Variant	Clarity	Content	Aesthetics	Relevance	Overall
Baseline (Vanilla Bagel)	91.3	64.2	84.0	44.9	50.9
Pipeline Ablations					
w/o Text-Search	90.7 (-0.6)	70.9 (+6.7)	84.3 (+0.3)	64.6 (+19.7)	65.4 (+14.5)
w/o Image-Search	**92.1** (+0.8)	73.1 (+8.9)	85.0 (+1.0)	50.8 (+5.9)	56.2 (+5.3)
w/o Recaption	83.0 (-8.3)	69.0 (+4.8)	74.5 (-9.5)	60.2 (+15.3)	62.9 (+12.0)
Constraint Ablations					
Recaption w/o Identity-preserving	91.5 (+0.2)	72.6 (+8.4)	83.4 (-0.6)	65.9 (+21.0)	67.7 (+16.8)
Recaption w/o Scene-compositional	90.7 (-0.6)	70.8 (+6.6)	80.7 (-3.3)	68.6 (+23.7)	68.2 (+17.3)
Recaption Architecture Ablations					
Recaption w/o VAE	90.9 (-0.4)	74.3 (+10.1)	84.5 (+0.5)	70.8 (+25.9)	71.2 (+20.3)
Recaption w/o ViT	88.6 (-2.7)	68.4 (+4.2)	81.1 (-2.9)	58.7 (+13.8)	61.4 (+10.5)
Unify-Agent (Full)	91.2 (-0.1)	**75.2** (+11.0)	**85.2** (+1.2)	**72.4** (+27.5)	**73.2** (+22.3)

VAE latents mainly preserve appearance details (color, texture, material) that refine but cannot replace semantic grounding. Empirically, removing ViT causes a substantial drop in Relevance and Overall, while removing VAE yields a noticeably smaller drop—consistent with this division of roles.

We therefore draw a deliberately narrow conclusion: *joint access to ViT [14] and VAE [23] features within a single backbone* benefits recaption-driven world-grounded synthesis, with ViT contributing the larger share. We do not claim that generative priors broadly improve multimodal understanding; verifying such a stronger claim would require controlled comparisons between models with and without generative pretraining, which we leave to future work.

7 Limitations and Future Work

Our work still has several limitations. Current open-source unified multimodal models remain substantially weaker than the strongest closed-source systems; for example, Bagel still has limited long-context capability and can support only a relatively small number of images within a single context, which constrains more complex agent behaviors. Moreover, although we demonstrate strong results on IP- and concept-centric world-grounded synthesis, our current pipeline is still limited to a relatively shallow one-pass workflow, rather than more general iterative agent behaviors such as interleaved text-image search, reflection, and replanning, which are crucial for harder open-world tasks such as travel planning or academic report generation. At the same time, our results provide encouraging evidence for the feasibility of end-to-end unified agent training and highlight the advantage of unified multimodal models, where generation and understanding can mutually reinforce each other. In future work, we plan to validate these findings on stronger unified backbones and extend the framework toward more capable multimodal agents that can support longer-horizon planning, repeated search, and adaptive reasoning in more complex real-world settings.

8 Conclusion

We present **Unify-Agent**, a unified multimodal agent that reframes image generation from a closed-book prompt-to-image mapping into an inference-time process of cognitive gap detection, multimodal evidence acquisition, grounded recaptioning, and image synthesis. By explicitly resolving missing world knowledge through retrieved textual and visual evidence, our framework enables faithful generation of rare, long-tail, and knowledge-intensive concepts. Extensive experiments demonstrate that Unify-Agent consistently improves over its base unified model (Bagel) across diverse benchmarks and real-world image generation tasks. More broadly, our findings suggest that world-grounded image synthesis is a distinct multimodal reasoning problem requiring tightly coupled understanding, acting, and generation, motivating future work on more reliable open-world generative agents.

Acknowledgments and Disclosure of Funding

This work was supported by Tencent Hunyuan, which provided the computational resources and infrastructure used for all training, data construction, and evaluation reported in this paper. We thank Tencent Hunyuan for this support. The authors received no other third-party support in connection with this work during the last 36 months.

Competing interests. A subset of the authors are affiliated with Tencent Hunyuan, which develops and releases multimodal foundation models. Beyond this affiliation and the support described above, the authors declare no financial relationships outside the submitted work that could be perceived as influencing it.

References

[1] Wikipedia. https://www.wikipedia.org/.

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

[3] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

[4] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, Tiankai Hang, Duojun Huang, Jie Jiang, Zhengkai Jiang, Weijie Kong, Changlin Li, Donghao Li, Junzhe Li, Xin Li, Yang Li, Zhenxi Li, Zhimin Li, Jiaxin Lin, Linus, Lucaz Liu, Shu Liu, Songtao Liu, Yu Liu, Yuhong Liu, Yanxin Long, Fanbin Lu, Qinglin Lu, Yuyang Peng, Yuanbo Peng, Xiangwei Shen, Yixuan Shi, Jiale Tao, Yangyu Tao, Qi Tian, Pengfei Wan, Chunyu Wang, Kai Wang, Lei Wang, Linqing Wang, Lucas Wang, Qixun Wang, Weiyan Wang, Hao Wen, Bing Wu, Jianbing Wu, Yue Wu, Senhao Xie, Fang Yang, Miles Yang, Xiaofeng Yang, Xuan Yang, Zhantao Yang, Jingmiao Yu, Zheng Yuan, Chao Zhang, Jian-Wei Zhang, Peizhen Zhang, Shi-Xue Zhang, Tao Zhang, Weigang Zhang, Yepeng Zhang, Yingfang Zhang, Zihao Zhang, Zijian Zhang, Penghao Zhao, Zhiyuan Zhao, Xuefei Zhe, Jianchen Zhu, and Zhao Zhong. Hunyuanimage 3.0 technical report, 2026.

[5] Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19396–19405, 2025.

[6] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

[7] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[8] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

[9] 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.

[10] Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. A survey of multimodal hallucination evaluation and detection, 2026.

[11] 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.

[12] 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.

[13] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[15] 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.

[16] Google. Nano banana 2: Combining pro capabilities with lightning-fast speed. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2026.

[17] 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.

[18] Hsin-Ping Huang, Xinyi Wang, Yonatan Bitton, Hagai Taitelbaum, Gaurav Singh Tomar, Ming-Wei Chang, Xuhui Jia, Kelvin CK Chan, Hexiang Hu, Yu-Chuan Su, et al. Kitten: A knowledge-intensive evaluation of image generation on visual entities. *arXiv preprint arXiv:2410.11824*, 2024.

[19] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[20] Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, Junbo Qiao, Yue Guo, Yao Hu, Zhenfei Yin, Philip Torr, Yu Cheng, Wanli Ouyang, and Shaohui Lin. Interleaving reasoning for better text-to-image generation, 2025.

[21] Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, et al. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. *arXiv preprint arXiv:2601.22060*, 2026.

[22] Ziwei Huang, Wanggui He, Quanyu Long, Yandi Wang, Haoyuan Li, Zhelun Yu, Fangxun Shu, Long Chan, Hao Jiang, Fei Wu, et al. T2i-factualbench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. *arXiv preprint arXiv:2412.04300*, 2024.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[24] Black Forest Labs. Flux.1 [dev]: Text-to-image generation model. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2025.

[25] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[26] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

[27] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

[28] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models, 2025.

[29] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization, 2024.

[30] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv: 2209.03003*, 2022.

[31] Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. Realrag: Retrieval-augmented realistic image generation via self-reflective contrastive learning, 2025.

[32] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a very large multilingual semantic network. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[33] 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.

[34] OpenAI. Gpt-image-1.5 model documentation. https://developers.openai.com/api/docs/models/gpt-image-1.5, 2025.

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[38] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4695–4703, 2024.

[39] Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. Technical report, Technical report, Bytedance, 2025. URL https://lf3-static. bytednsdoc. com ....

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

[41] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. 2023.

[42] Moo Hyun Son, Jintaek Oh, Sun Bin Mun, Jaechul Roh, and Sehyun Choi. World-to-image: Grounding text-to-image generation with agent-driven world knowledge. *arXiv preprint arXiv:2510.04201*, 2025.

[43] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024.

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

[45] Gemini Team. Gemini: A family of highly capable multimodal models, 2025.

[46] OpenAI Team. Gpt-4o system card, 2024.

[47] Seed Team. Seed1.5-vl technical report, 2025.

[48] Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

[49] Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Li Donghao, Hang Tiankai, Li Zhenxi, Jiale Tao, QiXun Wang, Ruihuang Li, Comi Chen, Xin Li, Mingrui Wu, Xinchi Deng, Shuyang Gu, Chunyu Wang, and Qinglin Lu. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*, 2025.

[50] 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.

[51] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37:128374–128395, 2024.

[52] 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, 2025.

[53] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models, 2023.

[54] 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*, pages 12966–12977, 2025.

[55] 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.

[56] Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions, 2023.

[57] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Icml*, volume 3, page 7, 2024.

[58] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023.

[59] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. 2022.

[60] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation, 2025.

[61] Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, et al. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026.

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

Appendix

Appendix Contents

A Additional Benchmark Results **16**

B Implementation Details **19**

C Attention Masking Strategy **20**

D Evaluation Protocols **20**

E More Details about FactIP benchmark **23**

E.1 Construction 23

E.2 Results 23

E.3 Human Validation of FactIP Evaluation 23

F Prompt **25**

G Image Generation Showcases **33**

H FactIP Benchmark Evaluation Showcases **39**

A Additional Benchmark Results

To complement the main-paper results, this section provides the full quantitative tables and per-category discussion for FactIP, WiSE, KiTTEN, and T2I-FactBench, which were summarized in the main text due to space constraints.

FactIP (per-subtask). Tab. 3 reports the per-subtask overall scores on the FactIP benchmark, with all results evaluated on the representative FactIP-Mini split. The fine-grained breakdown reveals that, within the unified-MLLM group, Unify-Agent achieves the top Overall score (73.2) and dominates nearly every subcategory across Character (Celebrity, Animation, Game, Comic, Mythology, Mascot), Object (Animal, Food, Art, Toy), and Scene (Landmark, Festival), substantially outperforming Bagel and Bagel-CoT and rivalling several commercial systems. These results emphasize the necessity of advanced multimodal reasoning and extensive factual grounding to accurately render domain-specific visual characteristics, and underscore the efficacy of our unified agentic design in handling complex, knowledge-intensive generation tasks. A category-wise visual comparison and the construction details of the benchmark are deferred to the dedicated FactIP section below.

WiSE. Tab. 4 reports per-dimension WiScore on the WiSE benchmark, which evaluates world-knowledge-grounded generation across six knowledge dimensions: Cultural, Time, Space, Biology, Physics, and Chemistry. Unify-Agent achieves an Overall WiScore of 0.77, establishing a new state of the art among unified MLLMs and surpassing the strong BAGEL+CoT baseline (0.70). It attains the best results on Cultural (0.82), Time (0.75), Biology (0.72), and Chemistry (0.70), and remains highly competitive on Space (0.74) and Physics (0.73). These results indicate that the agentic search and grounded recaptioning pipeline meaningfully improves factual alignment across heterogeneous knowledge domains rather than overfitting to a single category.

KiTTEN. Tab. 5 reports per-category results on the KiTTEN benchmark, which evaluates fine-grained generation along two complementary axes: *text alignment* and *entity alignment*, across eight categories including Aircraft, Vehicle, Cuisine, Flower, Insect, Landmark, Plant, and Sport. Unify-Agent attains an overall score of 4.08, establishing a new state of the art across both generation-only and unified MLLM approaches and surpassing the strong Imagen-3 baseline (3.50). It also achieves the highest overall text alignment (4.22) and entity alignment (3.93) scores, while delivering balanced



Figure 5: High-quality samples from our **Unify-Agent**, highlighting its excellence in unified multi-image generation and agentic search enhanced world knowledge integration. It delivers strong cross-image consistency, broad stylistic versatility, and more faithful, knowledge-grounded visual generation across diverse concepts and scenarios—even for up-to-date real-world queries, such as generating images of the top three finishers of the 2026 Chinese Grand Prix in Shanghai.

Table 3: Per-subtask overall scores on the FactIP benchmark. **Character** includes celebrity, animation, game, comic, mythology, and mascot prompts; **Object** includes animals/plants, food, cultural relic/art, and toy prompts; **Scene** includes landmark and celebration prompts. denotes the best score and the second-best within each group.

Model	Character						Object				Scene		Overall
	Celebrity	Animation	Game	Comic	Mythology	Mascot	Animal	Food	Art	Toy	Landmark	Festival	
Commercial Models													
GPT-Image-1	77.0	63.1	66.4	60.3	95.7	64.7	**89.0**	88.6	40.9	52.2	76.1	94.2	69.3
GPT-Image-1.5	77.7	64.0	64.1	61.9	93.0	69.4	88.1	92.1	42.8	48.9	76.0	95.4	69.9
Seedream-4.5	86.0	75.6	82.3	80.8	88.9	88.1	77.0	92.4	64.8	83.0	83.5	88.9	82.0
Seedream-4	86.8	76.1	85.3	81.0	88.9	85.2	80.1	91.4	70.2	83.8	85.8	87.5	83.0
Seedream-5	87.5	**85.0**	**89.5**	**86.8**	87.8	**90.9**	81.5	**96.0**	77.5	89.5	**88.8**	82.8	87.3
Qwen-Image-2.0-Pro	68.8	64.5	71.9	64.6	80.5	74.3	75.5	82.7	65.0	79.3	80.2	84.0	71.1
Nano Banana	48.8	53.2	57.8	49.0	70.0	65.5	49.4	64.2	85.2	69.1	59.4	56.7	56.6
Nano Banana-Pro	75.0	61.4	62.9	54.6	**97.0**	63.6	87.0	89.9	43.8	46.3	72.3	90.3	66.7
Nano Banana-2	**92.5**	82.6	89.0	83.7	95.2	89.1	88.1	94.8	**86.9**	**94.8**	88.3	**96.4**	**88.5**
Generation Only													
Pixel-Art-XL	24.7	5.6	12.1	12.1	19.8	5.8	25.9	1.4	2.2	5.0	6.8	26.1	12.1
FLUX.1-schnell	26.0	16.8	25.9	30.1	50.0	17.4	49.6	11.8	12.1	18.9	19.7	54.9	24.0
SD-3-medium	33.7	17.8	23.8	31.5	49.1	18.7	49.0	13.1	11.3	16.6	19.2	58.9	25.7
SDXL-base-0.9	38.2	18.7	27.5	29.9	40.8	15.3	43.2	11.5	11.8	17.4	18.7	60.1	26.5
SD-3.5-medium	33.4	19.7	27.2	32.9	50.7	16.2	47.0	8.8	11.1	16.8	21.2	64.3	26.5
SD-3.5-large	35.6	19.8	30.4	31.3	45.7	24.6	42.8	13.7	12.3	17.8	21.0	66.7	27.5
FLUX.1-dev	36.1	21.4	28.1	31.9	57.7	22.6	50.9	18.0	15.8	20.9	23.2	58.8	28.9
Z-Image	49.6	52.7	54.7	51.3	67.6	58.1	**77.3**	**75.4**	32.1	34.9	62.4	78.2	54.2
FLUX.2-dev	**54.2**	52.9	**55.8**	**52.2**	**90.4**	**62.9**	75.3	74.0	33.7	**38.6**	61.9	**81.9**	**56.3**
Qwen-Image	52.5	**56.3**	52.2	50.8	79.0	62.5	68.8	69.2	**33.9**	37.9	**65.8**	73.3	55.4
Unified MLLM													
Janus-1.3B	16.4	18.4	23.5	26.2	25.7	13.3	38.6	13.8	3.8	9.6	20.2	33.7	19.3
Janus-Pro	23.0	31.6	33.1	35.5	61.1	27.3	43.5	39.1	9.3	16.7	30.3	45.7	30.3
Emu3	30.6	28.6	31.7	33.8	36.5	21.3	45.6	26.1	15.5	21.3	30.9	49.0	30.0
Emu3.5	59.2	54.1	58.7	52.7	76.0	57.7	74.2	76.2	28.7	34.1	61.3	**87.0**	57.2
Hunyuan-Image-3.0	47.4	53.2	47.6	48.6	**82.5**	54.5	70.5	78.8	33.1	36.0	64.2	81.5	53.4
Echo4o	44.9	48.1	43.3	43.5	66.5	52.1	60.5	62.2	27.7	30.3	55.1	70.0	47.3
Bagel	46.1	53.7	46.5	48.2	63.3	54.3	66.6	76.5	31.6	33.5	53.5	75.2	50.9
Bagel-CoT	44.7	48.0	44.6	46.0	62.0	43.7	58.2	63.0	27.0	30.8	50.5	72.8	47.0
Unify-Agent (Ours)	**71.9**	**71.7**	**68.0**	**69.0**	76.5	**75.7**	**75.5**	**80.2**	**65.3**	**76.3**	**70.9**	79.2	**73.2**

Table 4: Performance comparison on the WiSE benchmark across different knowledge dimensions (WiScore). denotes the best score and the second-best within each group.

Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
Commercial Models							
GPT-Image-1	0.81	0.71	0.89	0.83	0.79	0.74	0.80
Nano Banana-Pro	**0.89**	0.80	0.89	0.88	0.86	**0.85**	0.87
Nano Banana	**0.89**	**0.87**	**0.95**	**0.89**	**0.89**	0.79	**0.89**
Generation Only							
SD-v1-5	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SD-2-1	0.30	0.38	0.35	0.33	0.34	0.21	0.32
FLUX.1-schnell	0.39	0.44	0.50	0.31	0.44	0.26	0.40
SD-3-medium	0.42	0.44	0.48	0.39	0.47	0.29	0.42
SD-XL-base-0.9	0.43	0.48	0.47	0.44	0.45	0.27	0.43
SD-3.5-medium	0.43	0.50	0.52	0.41	0.53	0.33	0.45
SD-3.5-large	0.44	0.50	0.58	0.44	0.52	0.31	0.46
PixArt-Alpha	0.45	0.50	0.48	0.49	0.56	0.34	0.47
playground-v2.5	0.49	0.58	0.55	0.43	0.48	0.33	0.49
FLUX.1-dev	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Qwen-Image	**0.62**	**0.63**	**0.77**	**0.57**	**0.75**	**0.40**	**0.62**
Unified MLLM							
Janus-1.3B	0.16	0.26	0.35	0.28	0.30	0.14	0.23
Janus-Pro-1B	0.20	0.28	0.45	0.24	0.32	0.16	0.26
vila-u-7b-256	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Janus-Pro-7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Emu3	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Hunyuan-Image-3.0	0.58	0.57	0.70	0.56	0.63	0.31	0.57
BLIP3o-8B	0.49	0.51	0.63	0.54	0.63	0.37	0.52
MetaQuery-XL	0.56	0.55	0.62	0.49	0.63	0.41	0.55
UniWorld-V1	0.53	0.55	0.73	0.45	0.59	0.41	0.55
BAGEL	0.44	0.55	0.68	0.44	0.60	0.39	0.52
BAGEL+CoT	0.76	0.69	**0.75**	0.65	**0.75**	0.58	0.70
Unify-Agent (Ours)	**0.82**	**0.75**	0.74	**0.72**	0.73	**0.70**	**0.77**

top performance across nearly every sub-category. These results highlight the model's ability to simultaneously preserve intricate visual details and faithfully adhere to entity-specific constraints.

Table 5: KiTTEN benchmark results (scores range from 0 to 5). Each category shows Text alignment and Entity alignment scores. denotes the best score and the second-best within each group.

Model	Aircraft		Vehicle		Cuisine		Flower		Insect		Landmark		Plant		Sport		Overall		
	Text	Entity	Text	Entity	Text	Entity	Text	Entity	Text	Entity	Text	Entity	Text	Entity	Text	Entity	Text	Entity	All
Generation Only																			
SD	3.06	2.34	3.62	3.12	2.97	3.09	4.22	2.28	2.84	1.22	3.78	1.97	3.69	1.69	3.47	3.44	3.46	2.39	2.92
Flux	3.31	2.41	**4.31**	2.97	**4.28**	2.47	4.12	1.75	3.50	1.12	**3.94**	1.72	4.44	1.22	3.75	2.69	3.96	2.04	3.00
Custom-Diff	3.22	2.94	3.50	3.41	3.53	1.78	3.72	3.28	2.66	1.88	3.44	3.34	3.47	2.91	2.75	2.03	3.29	2.70	3.00
Imagen	3.16	2.81	3.78	3.16	3.75	3.25	4.12	2.72	2.88	1.12	3.53	2.41	3.94	1.59	3.75	3.19	3.61	2.53	3.07
DreamBooth	3.16	3.12	3.66	3.50	3.31	2.66	3.88	2.59	2.94	1.91	3.53	3.53	3.41	2.59	3.09	2.66	3.37	2.82	3.09
Instruct-Imagen	2.53	**3.88**	3.25	**4.06**	2.75	3.91	2.41	**4.03**	1.91	**3.22**	2.94	**4.09**	3.00	**3.31**	2.28	3.25	2.63	**3.72**	3.17
Imagen-3	**3.41**	3.19	4.25	3.34	4.25	**4.06**	**4.38**	3.03	**4.09**	1.31	3.88	2.25	**4.72**	1.69	**4.41**	**3.75**	**4.17**	2.83	**3.50**
Unified MLLM																			
Bagel	3.51	2.97	4.46	4.08	3.91	3.40	3.44	2.64	2.65	1.58	3.31	1.79	3.01	1.71	3.23	2.98	3.44	2.64	3.04
Bagel-CoT	3.67	2.94	4.42	3.98	4.10	4.01	3.56	3.12	2.98	2.12	3.41	2.11	3.31	2.26	3.84	3.63	3.66	2.02	2.84
Unify-Agent (Ours)	**4.03**	**3.59**	**4.54**	**4.09**	**4.36**	**4.13**	**4.28**	**4.16**	**4.17**	**3.97**	**3.99**	**3.76**	**4.46**	**3.81**	**3.95**	**3.92**	**4.22**	**3.93**	**4.08**

Table 6: Comparison on T2I-FactBench (SKCM, SKCI, and MKCC). denotes the best score and the second-best within each group.

Model	SKCM	SKCI			MKCC			
	Concept	Concept	Instantiation	All	Concept	Instantiation	Composition	All
Generation Only								
SD v1.5	40.5	52.9	53.9	53.4	37.6	13.4	15.1	22.0
Pixart	26.4	46.2	55.3	50.8	35.8	19.8	24.3	26.6
SD XL	45.8	59.9	65.8	62.8	51.7	28.0	35.4	38.4
Playground	45.6	66.1	62.5	64.3	53.8	35.4	44.8	44.7
Flux.1-dev	35.6	54.9	58.0	56.5	56.9	54.1	63.8	58.3
SD-3.5	46.2	64.6	71.2	67.9	68.9	59.2	75.5	67.9
DALLE-3	**55.5**	**72.4**	**88.5**	**80.5**	**71.3**	**70.2**	**85.6**	**75.7**
Unified MLLM								
Bagel	31.8	52.3	64.2	58.2	68.1	57.6	73.2	66.3
Bagel-CoT	34.6	52.6	64.5	58.5	67.2	60.3	**77.1**	68.2
Unify-Agent (Ours)	**69.2**	**75.3**	**79.6**	**77.4**	**76.1**	**64.8**	73.6	**71.5**

T2I-FactBench. Tab. 6 presents results on T2I-FactBench, a three-tiered benchmark that measures factual accuracy of T2I models along Single Knowledge Concept Memorization (SKCM), Single Knowledge Concept Instantiation (SKCI), and Multiple Concept Composition with Interaction (MKCC). Within the unified MLLM group, Unify-Agent achieves the top overall scores on SKCI (77.4) and MKCC (71.5), and reaches an SKCM concept score of 69.2 that compares favorably even against commercial models such as DALLE-3 (55.5). This progression indicates that Unify-Agent captures the intrinsic visual attributes of single concepts (SKCM), adapts them to varied conditions such as differing actions or scenes (SKCI), and handles the explicit semantic relationships required for multi-concept interactions (MKCC).

B Implementation Details

The Supervised Fine-Tuning (SFT) is conducted on our self-constructed interleaved image-and-text dataset with 64 NVIDIA H20 GPUs. The entire fine-tuning process lasted for approximately 10 days, completing a total of 10k gradient steps. For the training objective, our model is optimized using a combined loss formulation that jointly updates the language model, the ViT encoder, and the multimodal connectors, while keeping the continuous generation VAE weights strictly frozen. Specifically, we apply a standard Cross-Entropy (CE) loss for autoregressive text generation and visual token prediction, alongside a Mean Squared Error (MSE) loss on the image-reconstruction latents for the visual generation branch. Both the CE and MSE losses are equally scaled with a weight of 1.0. Furthermore, to improve the model's instruction-following capability in multimodal contexts, we enable CE loss reweighting, which specifically increases the penalty weight on designated special tokens. To stabilize the optimization over ultra-long interleaved sequences, we apply a constant learning rate scheduler with a peak learning rate of 5×10^{-5} after 500 linear warmup steps, and enforce gradient clipping with a maximum L2 norm of 5.0. A comprehensive list of hyperparameters for the SFT stage is summarized in Table 7.

Table 7: Hyperparameters for the Supervised Fine-Tuning (SFT) stage.

Hyperparameter	**Value**
Model & Architecture Setup	
Base LLM	Bagel-14B
Vision Encoder (ViT)	SigLIP-SO400M-14 (NaViT)
Visual Generation VAE	FLUX VAE (Frozen)
ViT Patch Size	14
VAE Latent Patch Size	2
Tied Word Embeddings	False
Data & Sequence Length	
Max Tokens per Sample	40,240
Expected Tokens per Batch	40,240
Max Packed Tokens (Hard Limit)	41,520
Vit Transform	(378, 980)
Vae Transform	(512, 1024)
Optimization & Objective	
Optimizer	AdamW
Adam $(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.95, 10^{-15})$
Learning Rate	5×10^{-5}
LR Scheduler	Constant
Warmup Steps	500
Total Training Steps	10,000
Max Gradient Norm	5.0
Cross-Entropy (CE) Weight	0.25
Reconstruction (MSE) Weight	1.0
CE Loss Special Token Reweighting	True
Special Token CE Weight	3.0
Hardware & Distributed Setup	
Hardware	$64\times$ NVIDIA H20 GPUs
Training Duration	$\sim$ 10 Days
Parallel Strategy	FSDP (`HYBRID_SHARD`)

C Attention Masking Strategy

We implement a hybrid attention masking strategy tailored for interleaved agentic data to balance sequential logic with precise visual grounding. For the textual components of the sequence, including agentic dialogs and reasoning steps, a standard causal masking is applied to preserve the temporal and logical flow of the conversation. In contrast, retrieved reference images—represented by a composite of VAE and ViT tokens—employ a full attention mechanism, allowing these tokens to interact globally to facilitate holistic visual feature extraction. Notably, during the subsequent image-generation phase, the latent flow-matching tokens (i.e., VAE noisy tokens) are specifically restricted to attending exclusively to the preceding recaption tokens and the retrieved reference image tokens. By selectively masking out irrelevant textual reasoning traces and historical dialogs at this stage, we prevent large segments of high-noise text from interfering with the synthesis process, thereby ensuring that the generated visual content remains precisely grounded in the condensed recaptioning and the provided visual evidence. See Fig. 6 for details of our attention masking strategy.

D Evaluation Protocols

To ensure rigorous and standardized evaluation across benchmarks with distinct task formulations and assessment objectives, we adhere to the official evaluation protocols of each benchmark whenever possible. For all existing benchmarks, we preserve their original MLLM-based evaluation settings, including the designated expert evaluators and scoring schemes, to maintain consistency with prior literature and improve the fairness and comparability of our results.

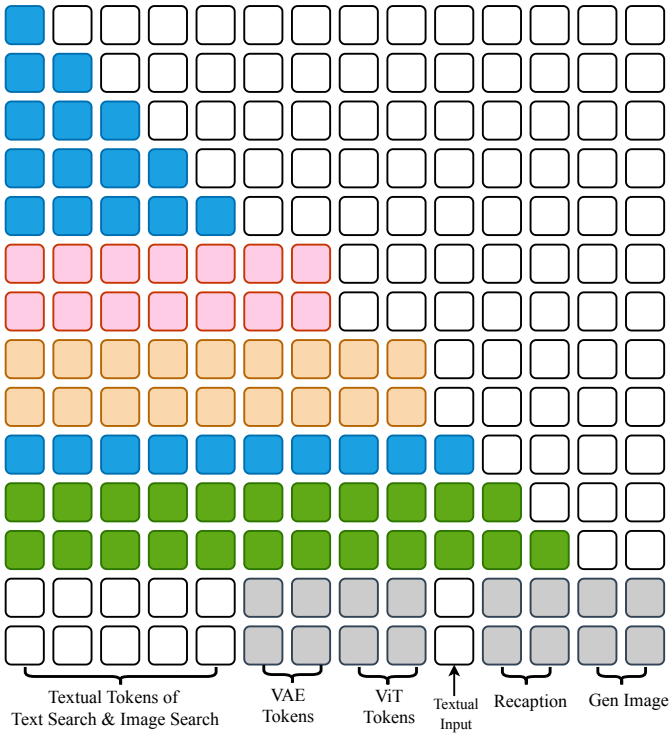


Figure 6: **Hybrid attention masking strategy for agentic image generation.** The matrix shows causal attention for reasoning traces and textual dialog (blue) and full attention for reference visual-textual evidence (pink/orange/green). Notably, the generation phase (grey) employs a restricted mask that filters out historical reasoning noise, forcing the latent flow-matching tokens to attend only to relevant reference images and recaptions for high-fidelity synthesis.

WiSE Evaluation. For the **WiSE Benchmark**, we employ **GPT-4o** [46] as the expert evaluator to assess the alignment between generated images and world knowledge. Given the textual prompt and the model-generated result, the evaluator is instructed to score the generation from three complementary aspects: *Consistency*, *Realism*, and *Aesthetic Quality*. Specifically, *Consistency* evaluates the accuracy and completeness with which the generated image reflects the prompt; *Realism* assesses the physical plausibility, including adherence to physical laws and accurate material representation; and *Aesthetic Quality* measures the overall artistic appeal, including composition and color harmony. Each dimension is evaluated on a discrete scale from 0 to 2 (0 for Rejected, 1 for Conditional, and 2 for Exemplary).

To obtain a unified measure of knowledge-image alignment, we define the overall metric, **WiScore**, as a weighted combination of the three dimensions:

$$\text{WiScore} = \alpha_1 \cdot \textbf{Consistency} + \alpha_2 \cdot \textbf{Realism} + \alpha_3 \cdot \textbf{Aesthetic Quality}, \tag{6}$$

$$\text{subject to} \quad \alpha_1 + \alpha_2 + \alpha_3 = 1. \tag{7}$$

In our experiments, we set $\alpha_1 = 0.7$, $\alpha_2 = 0.2$, and $\alpha_3 = 0.1$. This configuration explicitly prioritizes *Consistency*, reflecting that accurately representing the intended objects and relationships grounded in world knowledge is the most critical criterion. Meanwhile, *Realism* and *Aesthetic Quality* are retained as auxiliary dimensions to ensure overall visual excellence without overshadowing the core knowledge assessment.

KiTTEN Evaluation. For the **KiTTEN Benchmark**, we utilize **GPT-4o** [46] as the expert evaluator to systematically assess the fine-grained visual fidelity of generated entities. Given the textual prompt, ground-truth reference images of the target entity, and the model-generated result, the evaluation framework decouples the assessment into two primary dimensions: *Text Alignment* and *Entity Alignment*. Specifically, *Text Alignment* evaluates the degree to which the generated image faithfully adheres to the prompt instructions beyond just the entity; and *Entity Alignment* assesses how well

the generated image captures precise visual details and realistically resembles the target entity based on the provided reference images. Each dimension is independently scored on an integer scale from 1 to 5, where 1 indicates no alignment and 5 denotes complete faithfulness. This decoupled design deliberately avoids fusing the scores into a single metric, allowing for a nuanced analysis of the trade-offs between preserving specific entity knowledge and strictly following complex textual constraints.

T2I-FactBench Evaluation. For the T2I-FactBench, we use a multi-round VQA protocol with GPT-4o across three levels: SKCM, SKCI, and MKCC. The evaluation consists of three progressive stages.

In the first round, **Concept Factuality** assesses the generated knowledge concept c_i in image I_i against its reference I_i^R across four dimensions: Shape (S), Color (C), Texture (T), and Feature Details (F). For each dimension, GPT-4o assigns a binary score (0 or 1). To account for multiple concepts at the MKCC level, the score is defined as:

$$\text{Concept Factuality} = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{S_{ij} + C_{ij} + T_{ij} + F_{ij}}{4} \right), \tag{8}$$

where N_i denotes the number of concepts within I_i, and $S_{ij}, C_{ij}, T_{ij}, F_{ij}$ represent binary scores for the j-th concept.

The second round, **Instantiation Completeness**, verifies whether concept c is present and if the instantiation phrase p is successfully realized. A score of 1 is assigned only if both conditions are satisfied.

Finally, for the MKCC level, a third round evaluates **Composition Factuality** across four dimensions: Seamless Transition (S_i), Visual Completeness (V_i), Authenticity (A_i), and Prompt Following (P_i). The composition factuality score is defined as:

$$\text{Composition Factuality} = \frac{S_i + V_i + A_i + P_i}{4}, \tag{9}$$

where $S_i, V_i, A_i, P_i \in \{0, 1\}$ represent scores for the respective dimensions for the i-th prompt.

FactIP Benchmark Evaluation. For our proposed **FactIP Benchmark**, we employ **Seed2.0** [39] as the expert evaluator to conduct a fine-grained and structured assessment of generated images and the evaluation prompt can be found in F. Given the textual prompt, two reference images (**GT1** and **GT2**), and the model-generated result (**AS**), the evaluator is instructed to jointly compare the generated image against both the prompt and the references, and to score it from four complementary aspects: *Clarity*, *Content*, *Aesthetics*, and *Relevance*. Specifically, *Clarity* evaluates image sharpness, visual cleanliness, and the richness of perceptual details; *Content* measures whether the generated image faithfully captures the key semantic elements and compositional requirements specified in the prompt; *Aesthetics* assesses the overall visual quality, including composition, lighting, color harmony, and stylistic appeal; and *Relevance* measures whether the generated image preserves the intended IP identity and remains consistent with the defining attributes inferred from the two reference images. Each dimension is scored on an integer scale from 0 to 10 by the expert evaluator. Following the evaluation protocol, the two references (**GT1** and **GT2**) are considered jointly, such that stable identity traits are emphasized while variations already existing across the references are not over-penalized.

To obtain a unified measure of generation quality, we define the overall score as a weighted combination of the four dimensions:

$$\text{Overall Score} = \alpha_1 \cdot \textbf{Clarity} + \alpha_2 \cdot \textbf{Content} + \alpha_3 \cdot \textbf{Aesthetics} + \alpha_4 \cdot \textbf{Relevance}, \tag{10}$$

$$\text{subject to} \quad \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1. \tag{11}$$

In our experiments, we set $\alpha_1 = 0.05$, $\alpha_2 = 0.10$, $\alpha_3 = 0.10$, and $\alpha_4 = 0.75$. This configuration explicitly prioritizes *Relevance*, reflecting that consistency with the actual IP identity is the most critical criterion in FactIP Benchmark. Such a design is also aligned with our emphasis on broad world knowledge utilization, since accurately preserving the target IP requires the model to recognize and faithfully express knowledge-intensive identity cues rather than merely producing visually plausible content. Meanwhile, *Clarity*, *Content*, and *Aesthetics* are retained as important auxiliary dimensions

to ensure that the final metric also captures perceptual quality and semantic completeness. Finally, the weighted overall score is multiplied by 10, so that the final score range is normalized to $[0, 100]$, where a higher score indicates better IP fidelity and overall generation quality. More Evaluations showcases can be found in H.

E More Details about FactIP benchmark

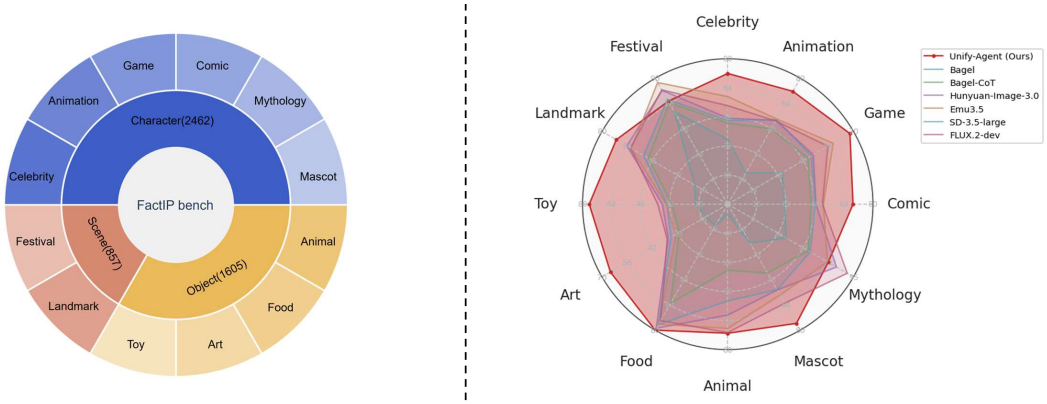


Figure 7: (a) Hierarchical category distribution of FactIP Bench, consisting of three major groups (Character, Scene, and Object) and 12 fine-grained subcategories. (b) Category-wise comparison of different methods on FactIP Bench, where the radar chart presents the overall scores across all subcategories.

E.1 Construction

To rigorously evaluate the capability of generative models in synthesizing high-fidelity intellectual property (IP) and culturally significant concepts, we construct the FactIP benchmark. As illustrated in Figure 7(a), the benchmark is meticulously organized into a hierarchical taxonomy. It encompasses three overarching domains—Character, Object, and Scene—which are further decomposed into fine-grained subcategories. This comprehensive structure ensures a diverse and representative evaluation of a model's factual knowledge grounding. For a detailed statistical breakdown and specific descriptions of each subcategory, please refer to Table 8. In total, the dataset comprises 2,462 meticulously curated prompts, challenging models to generate recognizable entities ranging from global celebrities and intricate art toys to iconic landmarks. To facilitate efficient and accessible model evaluation, we additionally construct a lightweight version, FactIP-Mini, by sampling 500 prompts from the full dataset while strictly preserving the original category distribution.

E.2 Results

We present a comprehensive performance analysis of various leading generative models on the FactIP benchmark, with all reported results evaluated on the representative FactIP-Mini dataset. Figure 7(b) provides a visual category-wise comparison, highlighting the relative strengths and weaknesses of different architectures across the primary evaluation dimensions. The radar chart succinctly captures the overarching performance trends, revealing that Unified MLLMs and state-of-the-art commercial models generally exhibit superior IP fidelity and factual consistency compared to earlier generation-only paradigms. A detailed, quantitative breakdown of the per-subtask overall scores is provided in Table 3 (reported earlier in Sec. A). These fine-grained results emphasize the necessity of advanced multimodal reasoning and extensive factual grounding to accurately render domain-specific visual characteristics. Notably, our proposed Unify-Agent demonstrates highly competitive and balanced performance across all taxonomic branches, underscoring the efficacy of its unified architectural design in handling complex, knowledge-intensive visual generation tasks.

E.3 Human Validation of FactIP Evaluation

To validate the reliability of our automatic FactIP evaluation, we conduct a human preference study on 200 randomly sampled prompts from FactIP-Mini, preserving the original category distribution

Category	Subcategory	Description	Num
CHARACTER	Animation	Animated characters, creatures, equipment, and iconic locations from anime and animated media.	438
	Comic	Characters and visual elements originating from comic books and manga series.	363
	Celebrity	Prominent figures across diverse domains, including scientists, political leaders, business executives, athletes, and entertainment personalities.	300
	Game	Video game characters, weapons, equipment, and other in-game visual elements.	272
	Mascot	Official mascots representing Olympic Games, regional events, and corporate brands.	77
	Mythology	Universally recognized mythological narratives and legendary figures, e.g., Kuafu Chasing the Sun.	50
OBJECT	Food	Cuisines, regional delicacies, desserts, and beverages with cultural significance.	316
	Cultural Relic / Art	National treasures, classical calligraphy, paintings, sculptures, and fine art pieces.	126
	Toy	Collectible figures, designer toys, and model kits with cultural relevance, e.g., Labubu.	123
	Animal / Plant	Individually notable animals and plants with distinct public recognition, e.g., Giant Panda Qizai.	50
SCENE	Landmark	Renowned scenic spots, architectural landmarks, monuments, and heritage sites.	297
	Festival / Celebration	Visual elements and symbols associated with well-known festivals and cultural celebrations.	50

Table 8: Detailed category statistics of our benchmark. The dataset contains three major categories and 12 fine-grained subcategories, totaling 2,462 prompts.

across Character, Object, and Scene. For each prompt, we compare Unify-Agent with representative baselines using the same generated images as in the main experiments. Each comparison is shown to three human annotators in a blind pairwise format, where annotators are given the original prompt, two reference images of the target IP, and two anonymized generated images in randomized order. Annotators are asked to select which image better preserves the target IP identity and follows the user instruction. Ties are allowed when both generations are considered comparably faithful.

Human preference protocol. All annotations are collected from trained professional annotators with prior experience in image quality and identity-preservation evaluation tasks. Before the formal study, annotators go through a short calibration session on a small held-out set of examples to align their understanding of identity-defining attributes and acceptable variations; calibration examples are excluded from the final 200-prompt evaluation. We focus the annotation on factual identity preservation and prompt faithfulness, which are the core dimensions of FactIP. Annotators are instructed to consider identity-defining visual attributes such as facial structure, character-specific appearance, object-defining details, landmark-specific cues, and culturally significant elements. At the same time, they are asked not to over-penalize acceptable variations in pose, lighting, background, or style, as long as the target identity remains recognizable and the user instruction is followed. Each pairwise comparison is independently judged by three annotators, and the final preference is determined by majority vote. We retain ties rather than forcing a binary choice, since multiple generations may be comparably faithful for open-ended image synthesis.

We report the win, tie, and lose rates of Unify-Agent against each baseline, as well as a tie-adjusted win rate:

$$\text{WinRate}_{\text{adj}} = \frac{\#\text{Win} + 0.5 \times \#\text{Tie}}{\#\text{Win} + \#\text{Tie} + \#\text{Lose}}.$$

Table 9: **Human preference evaluation on FactIP.** We report pairwise human preference between Unify-Agent and representative baselines on 200 randomly sampled prompts. Each comparison is judged by three annotators under a blind randomized order. Ties are retained when both generations are considered comparably faithful.

Comparison	Win (%)	Tie (%)	Lose (%)	Tie-adjusted Win (%)
Unify-Agent vs. Bagel	64.0	17.0	19.0	72.5
Unify-Agent vs. FLUX.1-dev	78.0	11.0	11.0	83.5
Unify-Agent vs. Hunyuan-Image-3.0	68.0	15.0	17.0	75.5
Unify-Agent vs. Emu3.5	65.0	17.0	18.0	73.5

Table 10: **Agreement between Seed2.0 and human preference on FactIP.** We compare Seed2.0 pairwise preferences with majority human votes on the same 200-prompt subset used in Table 9. Agreement is computed on non-tied human-majority cases.

Comparison	Seed2.0-Human Agreement (%)	Human Inter-Annotator Agreement (%)
Unify-Agent vs. Bagel	82.4	78.6
Unify-Agent vs. FLUX.1-dev	86.1	81.3
Unify-Agent vs. Hunyuan-Image-3.0	80.7	77.9
Unify-Agent vs. Emu3.5	81.5	78.2
Average	82.7	79.0

As shown in Table 9, human annotators consistently prefer Unify-Agent over strong open-source baselines. The preference is most pronounced against generation-only models such as FLUX.1-dev, and remains clear against unified MLLM baselines such as Bagel, Hunyuan-Image-3.0, and Emu3.5. These results confirm that the gains measured by automatic FactIP evaluation are aligned with human judgments on factual identity preservation and prompt faithfulness.

Agreement with MLLM-as-Judge evaluation. We further examine whether the automatic FactIP evaluator is aligned with human judgments. On the same 200-prompt subset, we compare Seed2.0 pairwise preferences with majority human votes. For each pairwise comparison, we convert the structured FactIP scores into a preference by selecting the image with the higher overall score. To avoid forcing a binary decision when two generations are nearly indistinguishable under the evaluator, we treat a comparison as a tie when the absolute score difference is below a fixed threshold of $\tau = 0.05$ on the normalized $[0, 1]$ FactIP score scale. This threshold is chosen a priori as a small fraction (5%) of the score range and is not tuned on the human-annotated subset. To verify that our conclusions are not sensitive to this choice, we additionally compute Seed2.0–human agreement under $\tau \in \{0.02, 0.05, 0.10\}$ and observe that the agreement varies by at most 2 percentage points across all four comparisons, indicating that the reported agreement is stable under reasonable choices of the tie threshold. Agreement is computed on examples with a non-tied human-majority decision. Human inter-annotator agreement is computed as the average pairwise agreement across the three annotators on each comparison, restricted to non-tied cases.

As shown in Table 10, Seed2.0 exhibits strong agreement with human preferences across all compared baselines, with an average agreement of 82.7%. The human inter-annotator agreement is 79.0% on average, indicating that factual identity preservation is a reasonably consistent criterion for human evaluation despite the open-ended nature of image generation. These results suggest that Seed2.0 provides a reasonable scalable proxy for human judgments on FactIP, while human evaluation remains the gold-standard validation.

F Prompt

System Prompt

SYSTEM_PROMPT

```
You are helping to build a high-quality visual generation
dataset.
Your task is to gather information and reference images for
creating detailed image descriptions.

Your Goal: Create a detailed recaption for "{image_prompt}"
```

```
about the IP "{ip_name}" ({country}).

Natural Workflow (think step by step):
1. First, search for background information about this
IP/character to understand who they are, their characteristics,
style, and context. This knowledge will help you craft more
accurate image search queries.
2. Then, search for reference images of this IP/character. Good
reference visuals are essential for the final detailed
description.
3. Finally, I will provide you with the downloaded reference
images, and you will generate a detailed <recaption> that
references "image_1" and "image_2" specifically.

Tool Call Format (IMPORTANT - use ONLY this format):
<tool_call>
{{"name": "tool_name", "arguments": {{"param1": "value1"}}}}
</tool_call>
Examples:
- Text search: <tool_call>{{"name": "text_search", "arguments":
  {{"q": "search query", "hl": "zh", "top_k": 5}}}}</tool_call>
- Image search: <tool_call>{{"name": "search_image",
"arguments":
  {{"q": "image query", "hl": "zh", "num": 8}}}}</tool_call>
```

TOOLS_DEFINITION

```
tools = [
    {
        "type": "function",
        "function": {
            "name": "text_search",
            "description": "Search the web for text information
            about a query. Use this to get background
            information
            about IPs, people, places, or topics.",
            "parameters": {
                "type": "object",
                "properties": {
                    "q": {
                        "type": "string",
                        "description": "Search query"
                    },
                    "hl": {
                        "type": "string",
                        "description": "Language code
                        (e.g., 'en', 'zh')",
                        "default": "en"
                    },
                    "top_k": {
                        "type": "integer",
                        "description": "Number of results
                        to return",
                        "default": 5
                    }
                },
                "required": ["q"]
            }
        }
    },
    {
        "type": "function",
```

```
            "function": {
                "name": "search_image",
                "description": "Search for images on the web. Use
                this to find relevant images about the IP or
                topic.",
                "parameters": {
                    "type": "object",
                    "properties": {
                        "q": {
                            "type": "string",
                            "description": "Search query for
                            images"
                        },
                        "location": {
                            "type": "string",
                            "description": "Location for search",
                            "default": "United States"
                        },
                        "hl": {
                            "type": "string",
                            "description": "Language code",
                            "default": "en"
                        },
                        "num": {
                            "type": "integer",
                            "description": "Number of images
                            to return",
                            "default": 8
                        }
                    },
                    "required": ["q"]
                }
            }
        }
    ]
```

Evluation Prompt

EVALUATION_PROMPT

```
You are an image evaluation assistant. Compare AS
(assistant-generated image) against GT1 and GT2 (ground-truth
images) given the Prompt.

Your task is to return exactly 6 fields: 5 integer scores (0-10)
and 1 rationale string.

Evaluate these 5 dimensions independently:

1. "clarity": image sharpness, absence of blur/artifacts/noise,
and richness of visible details.
2. "content_quality": faithfulness to the Prompt, subject
completeness, and semantic coherence.
3. "aesthetics": visual appeal, composition, lighting/color
harmony, and style consistency.
4. "text_relevance_ip": IP identity consistency. This means
whether AS preserves the same character/object/IP identity as
GT1/GT2 based on distinctive traits (e.g. face, hairstyle,
costume, colors, species/object-defining features). Do NOT
require exact matching of pose, background, camera angle, or
composition.

Important instructions:
```

```
- Use GT1 and GT2 jointly to infer the stable identity and
attributes of the IP.
- Do not penalize AS for differences that also vary between GT1
and GT2.
- Score each dimension independently before assigning the
overall score.
- All five score fields must be integers from 0 to 10.
- "rationale" must be a single short string with at least two
concrete evidence points, separated by semicolons.

You MUST respond with ONLY this exact JSON structure, with all
6 keys present and no extra keys:

{"clarity": 7, "content_quality": 8, "aesthetics": 7,
 "text_relevance_ip": 8,
 "rationale": "Evidence 1; Evidence 2"}

Do not use markdown fences.
Do not output any text before or after the JSON.
If uncertain, still output the best-effort JSON with all
required keys.
```

Judge Prompt

JUDGE_PROMPT

```
You are an expert Image Quality Assessor for an AI training
pipeline. Your task is to evaluate whether a downloaded image
is a high-quality visual reference for a specific Intellectual
Property (IP).

You must rate the image on a scale of 0 to 10 based on the
following strict rubrics:

### 1. IP Consistency (Critical)
- Pass: The image clearly depicts the specific character/object
requested.
- Fail (Score 0): The image shows a completely different
character, a landscape, a real person (cosplay) if the IP is
anime, or an unrelated object.

### 2. Layout & Composition
- Subject-Centric: The IP subject must be the main focus.
- Face Visibility: For characters, the face must be clearly
visible (preferably front-facing or 3/4 view). REJECT if the
face is tiny, distant, or back-facing.
- No Text-Heavy: REJECT images that are primarily movie posters
with large text overlays, book covers, or infographics where
the subject is obscured by text.
- No Collages: REJECT split-screens, manga panels, or multiple
images stitched together. Single-scene images only.

### 3. Visual Quality
- Clarity: The image must be sharp. REJECT if it is
  severely blurry, pixelated, or has heavy jpeg compression
  artifacts.

### 4. Watermarks & Obstructions
- Reject: Large, obstructive watermarks covering the face or
main body (e.g., full-screen stock photo watermarks).
- Accept: Small, unobtrusive logos in the corner are acceptable
but lower the score slightly.
```

```
    ### Scoring Guide:
    - 0: Wrong IP or completely unusable.
    - 1-5 (Reject): Correct IP but violates a major rule (Blurry,
    Huge Watermark, Text-Heavy, Collage, Tiny Face).
    - 6-7 (Borderline): Usable but not ideal (Side profile, small
    corner watermark, medium resolution).
    - 8-10 (Excellent): Perfect reference (High-res, clear
    front-facing, no text, clean).
```

JUDGE_Question

```
    Please evaluate the provided image for the IP: "{ip_name}".

    Assess the image based on the system rubrics.
    Return your response in JSON format ONLY with the following
    structure:
    {{
        "score": <int, 0-10>,
        "reason": "<string, a concise explanation of the score,
        mentioning any specific flaws like 'text-heavy',
        'watermark', or 'blurry'>",
        "is_text_heavy": <bool>,
        "has_watermark": <bool>
    }}
```

Prompt-Generation Prompt

PROMPT-GENERATION_PROMPT

```
    You are a specialized Image Generation Prompt Expert. Your
    mission is to generate a single, high-quality image generation
    prompt based on a specific IP (person/character). You must
    describe "Who is doing What in Which scene."

## 1. Core Logic & Language Rules (CRITICAL)

Step 1: Determine the IP's Nationality
* Case A: If the IP is Chinese (Mainland, Hong Kong, Taiwan,
  Macau):
    * Prompt Language: Must be Chinese.
    * Tag Language: Must be Chinese.
    * Language Code: `zh`
* Case B: If the IP is NOT Chinese (USA, UK, Japan, Korea,
  Europe, etc.):
    * Prompt Language: Must be English.
    * Tag Language: Must be English.
    * Language Code: `en`

## 2. Prompt Construction Requirements

1.  Content Elements:
    * Subject: The IP's name (and brief identity if needed for
    context).
    * Scene: A specific, realistic location fitting the IP's
    profession (e.g., Office, Studio, Stadium, Stage, Cafe,
    Street).
    * Action: A dynamic verb describing what they are doing (e.g.,
    Interviewing, Singing, coding, running, drinking coffee).
2.  Realism Constraint:
    * The scene and action must align with the IP's public persona
    or profession.
```

```
    * Do not hallucinate impossible scenarios unless the IP is a
    fictional fantasy character.
3.  Syntactic Diversity (Avoid Repetition):
    * Do not always use the structure "Name is doing X in Y".
    * Vary your sentence structures:
        * *Scene-first*: "In the [Scene], [Name] is [Action]..."
        * *Action-focused*: "[Name] is [Action] while located
          in [Scene]..."
        * *Descriptive*: "Surrounded by [Context], [Name] is
          [Action]..."

## 3. Output Format

You must output ONLY the XML tags below. Do not output markdown
code blocks (like ```xml), explanations, or conversational filler.

<Image_Prompt>The full descriptive sentence</Image_Prompt>
<Tag_Name>Category of the action (e.g., Interview, Speech,
Daily Life)</Tag_Name>
<Language>zh OR en</Language>

*Note on <Tag_Name>*: If Language is `zh`, the tag must be Chinese.
If Language is `en`, the tag must be English (e.g., "Hosting",
"Street Snap").

REMEMBER: Output ONLY the three XML tags above, nothing else.

## 4. Examples

Input: Robert Downey Jr. (American Actor)
Output:
<Image_Prompt>Sitting in a relaxed pose on a Hollywood talk
show set, Robert Downey Jr. is laughing while telling a
story</Image_Prompt>
<Tag_Name>Interview</Tag_Name>
<Language>en</Language>

Input: Gordon Ramsay (British Chef)
Output:
<Image_Prompt>Gordon Ramsay is carefully plating a gourmet dish
in a busy high-end restaurant kitchen</Image_Prompt>
<Tag_Name>Cooking</Tag_Name>
<Language>en</Language>
```

Recaption Prompt

RECAPTION_PROMPT

```
    You are a professional visual language reasoning assistant.
    Your task is to generate a reasoning process and a final
    detailed image description based on two reference images
    (image_1, image_2), the original instruction, and text
    search results.

Input Information
1. Reference Images: Explicitly labeled Reference Image 1 (refer to
as image_1) and Reference Image 2 (refer to as image_2).
2. Original Instruction: The user's short request.
3. Background Info: Text information from previous search steps.

Output Format
Strictly follow XML format:
<think>
[Deep reasoning here:
 1. Analyze visual features of image_1 and image_2.
 2. Combine with background info to plan the fusion.
 3. Explicitly state what comes from image_1 and what comes
    from image_2.]
</think>
<recaption>
[The final detailed image description, including
"Scene Description" and "Preservation Statement"]
</recaption>

Core Rules & Constraints
1. Reference Principle: You MUST strictly use "image_1" and
"image_2" to refer to the images. DO NOT use vague terms like "the
first image" or "reference picture".

2. Descriptive Style: The content of <Instruction> must be a
description of the final result (Descriptive), NOT an editing
command (Imperative).
   - BAD: "Please put the man from image_1 on the left..."
   - GOOD: "In this realistic outdoor portrait, the man from
   image_1 stands on the left..."

3. Preservation Statement: At the end of the description, you MUST
explicitly state what specifically is preserved from image_1 and
image_2.
   - Must include phrases like: "The final image completely
   preserves [features] from image_1..." and "The final image fully
   retains [features] from image_2...".
   - Facial Features Preservation (CRITICAL): If image_1 and/or
   image_2 contain identifiable persons, you MUST preserve their
   exact facial features as shown in the reference images. Include
   detailed descriptions such as:
   face shape, eyebrow shape, eye characteristics, facial
   expression, skin tone/complexion, hairstyle, and any distinctive
   facial features. Use the format: "Preserve the exact facial
   features of [person name/description] as shown in image_1 and
   image_2: [detailed facial feature description]. Maintain
   [their/his/her] [appearance/clothing/style] as referenced in
   both images."
```

```
4. Language Consistency (MANDATORY): The content of <recaption>
MUST be written entirely in the SAME language as the original
instruction. If the original instruction is in Chinese, ALL
descriptions in <recaption> must be in Chinese - NO English allowed.
If the original instruction is in English, use English throughout.
NEVER mix languages within the same description.

Start the task. Output <think> and <recaption>.
```

Summary Prompt

SUMMARY_PROMPT

```
Based on the following webpage content, provide a concise summary
that is relevant to the query: "{query}"

Webpage Title: {title}
Content:
{content[:2000]}

Please provide a focused summary (2-3 sentences maximum) that
directly addresses the query. Focus on the most relevant
information.

You MUST format your response using the following structure:
<think>
[Your thinking process about what information is most relevant to
the query]
</think>

<response>
[Your concise summary here - 2-3 sentences maximum]
</response>
```

G Image Generation Showcases

Case 1: Science - Copper Combustion This case details the generation of a scientifically accurate image for the knowledge-based prompt **The copper is burning, highlighting the color**. Recognizing the specific chemical knowledge required, the model first initiated a text search to verify the characteristic flame test reaction of copper, confirming it produces a distinctive green glow. To capture the nuanced visual dynamics, it then performed a targeted image search for "copper wire burning green flame." Through a rigorous image judgment process, the model filtered out low-quality or irrelevant references, selecting high-quality images that showcased both an intense emerald-green core and organic flame tendrils. During the refined recaptioning stage, the model synthesized these visual elements—integrating the luminous intensity of copper(II) compounds with the fluid motion of turquoise-green flames—while adding professional photographic descriptors, providing precise guidance for the generation, accurately depicting the mesmerizing chemical phenomenon. See Fig. 8 for details.

Case 2: Art Toy - DUDOO Tasty Afternoon Tea Series This case details the image generation process for the **DUDOO Tasty Afternoon Tea Series**. The model initially performed a round of text search to establish the general background of the series; however, as the first round yielded only high-level information about the product line, a second round was conducted to specifically probe for the visual identity and iconic character traits of "Dudu." This iterative search strategy compensated for the initial information gap, ensuring a comprehensive understanding of the IP's design elements before sourcing reference images. By integrating these refined character details with visual cues from the retrieved images, the model synthesized an IP-accurate recaption that vividly depicts the "popcorn bucket" theme and matte resin texture. See Fig. 9 for details.

Case 3: Celebrity - William Butler Yeats This case details the generation of an image for a prompt depicting the poet **William Butler Yeats** in his rustic stone cottage study. The model first initiated a text search to gather precise biographical and physical descriptions, confirming iconic details such as his signature pince-nez glasses and historical background. It then performed a targeted image search to retrieve authentic historical portraits, capturing his distinct facial structure across different ages. By integrating these verified visual characteristics—such as his angular bone structure and tousled hair—into a refined recaption, the model was able to situate the historically accurate figure within the specific atmospheric setting of a manuscript-writing scene. See Fig. 10 for details.

Case 4: Celebrity - Grigory Perelman This case details the image generation process for the legendary mathematician Grigori Perelman. The model initially performed a text search to establish his reclusive persona and historical context in St. Petersburg, providing the thematic foundation for the "sparsely furnished apartment" scene. It then conducted a targeted image search to capture his highly distinctive "bohemian scholarly" appearance, characterized by wild, untamed hair and a thick, flowing beard. Through the image judgement and refined recaptioning stages, the model meticulously analyzed his facial features across different ages—specifically noting the transition from dark to gray-streaked hair—and integrated these traits with atmospheric details like scattered papers and soft natural light. This strategy ensured the final generation achieved profound biographical accuracy while vividly portraying the intense focus of a brilliant mind. See Fig. 11 for details.

Case 5: Celebrity - Bruce Beutler This expert-focused case details the image generation process for **Nobel laureate Bruce Beutler in an advanced laboratory setting**. The model initially performed a text search to establish his professional identity as a Nobel Prize-winning immunologist and geneticist at UT Southwestern, ensuring the generated environment was contextually authentic. It then sourced reference portraits through image search and applied a rigorous image judgement process to filter for his most representative facial features, such as his receding gray-silver hair and distinctive mustache. By synthesizing these verified biographical details with the technical requirements of the scene in a refined recaption, the model achieved high accuracy in both persona and atmospheric realism. See Fig. 12 for details.

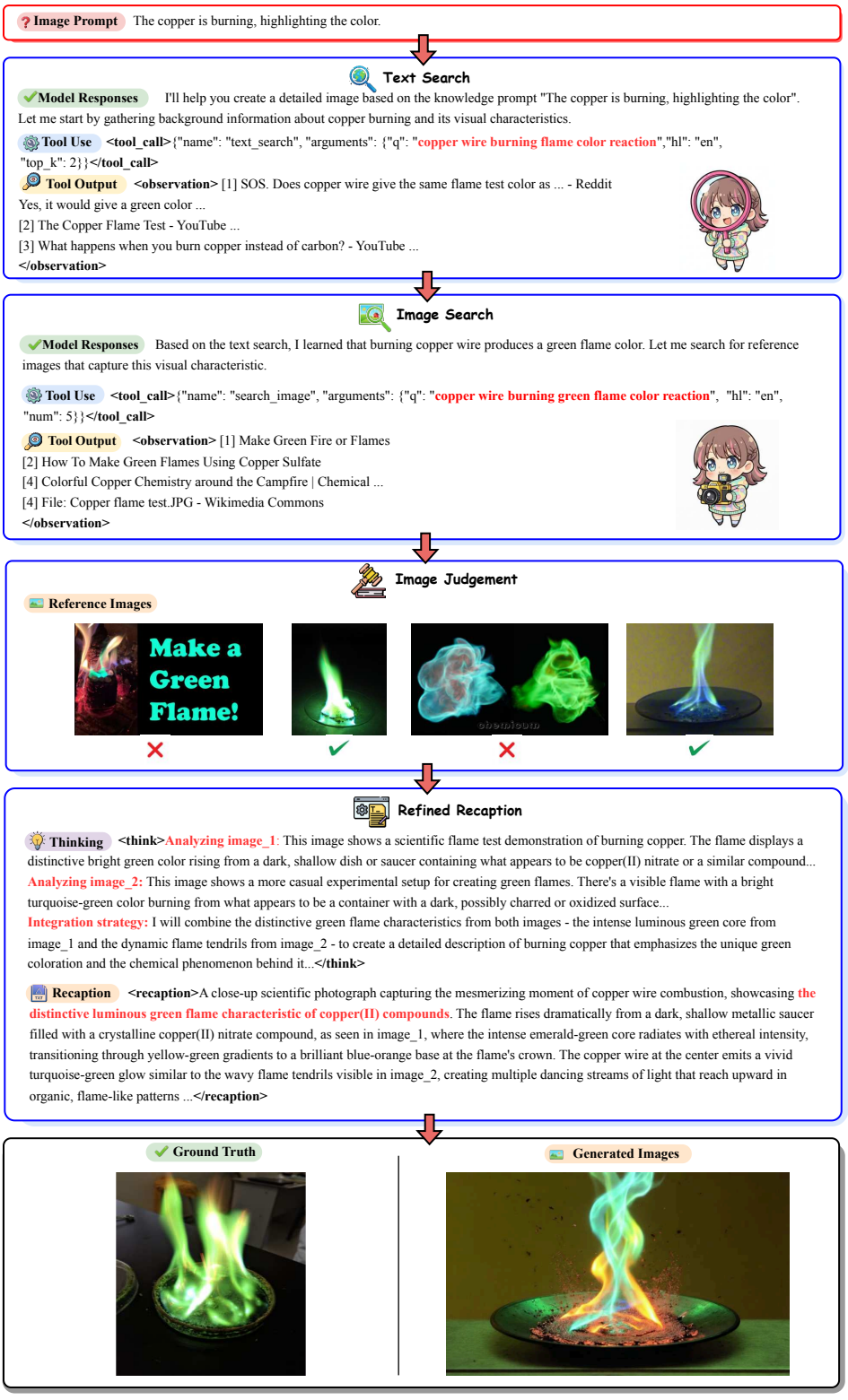


Figure 8: Image generated for the prompt: "The copper is burning, highlighting the color"



Figure 9: Image generated for the "DUDOO" art toy.

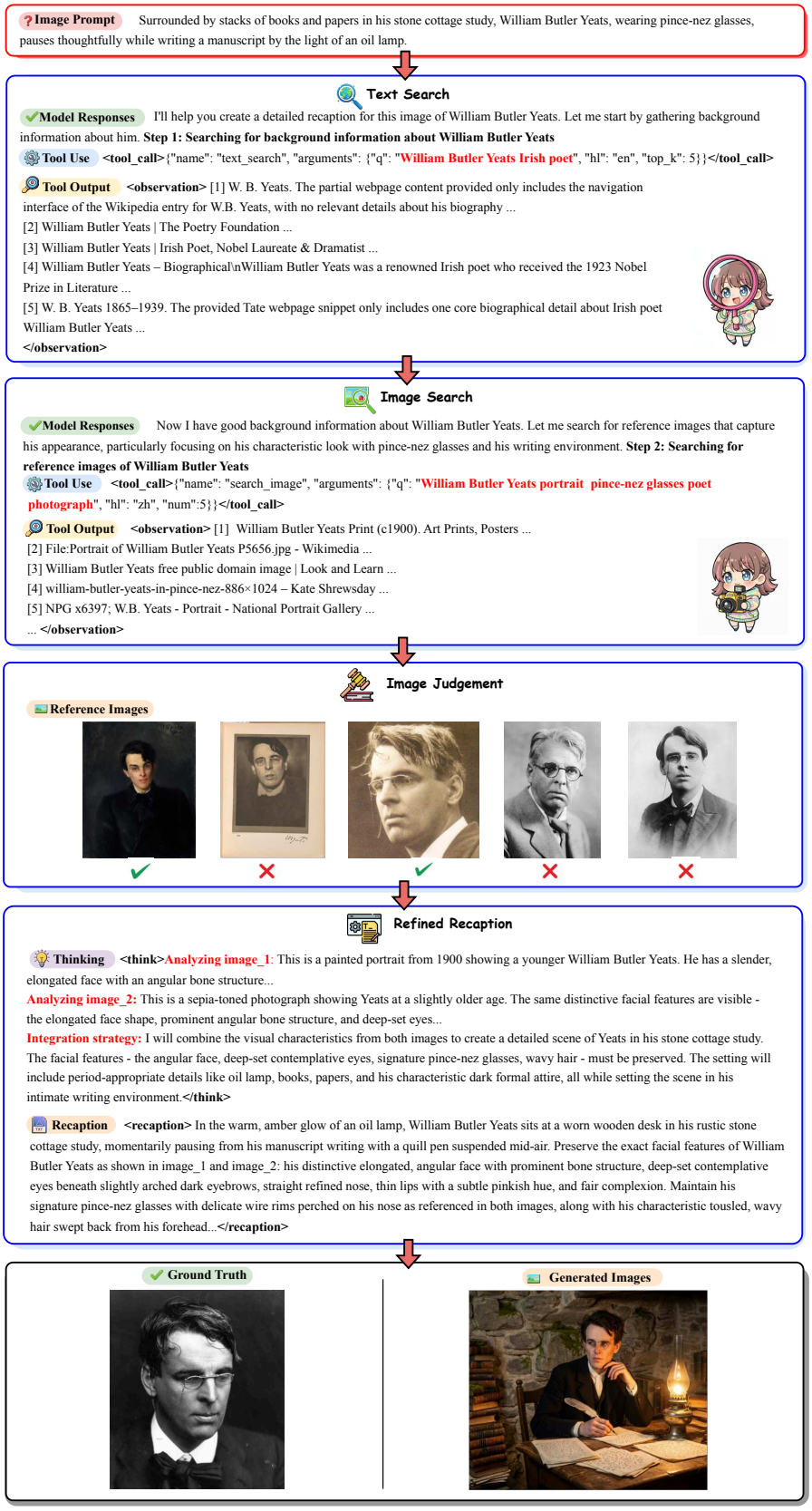


Figure 10: Image generated for William Butler Yeats writing a manuscript.

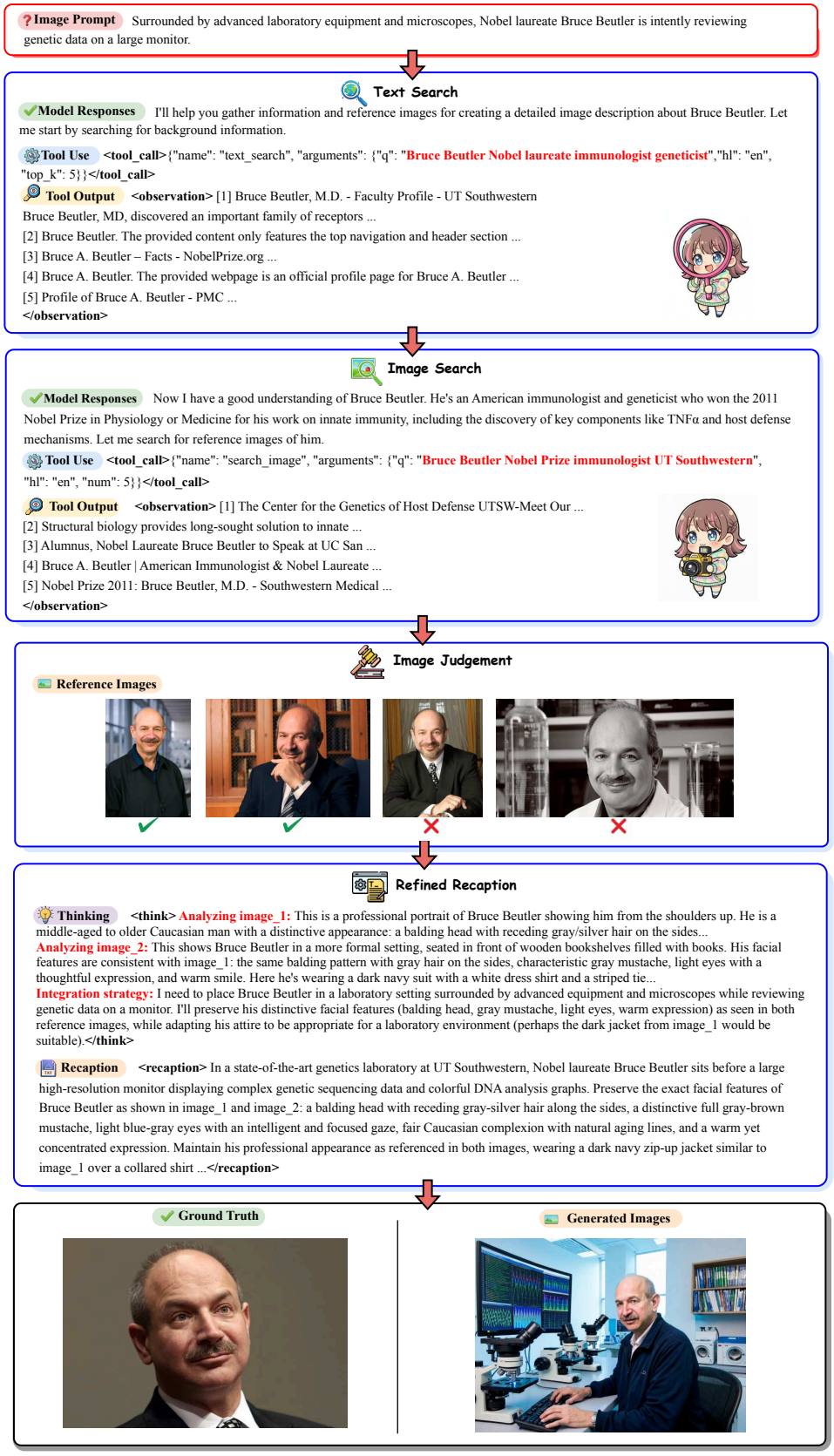


Figure 12: Image generated for Bruce Beutler reviewing data on a monitor.

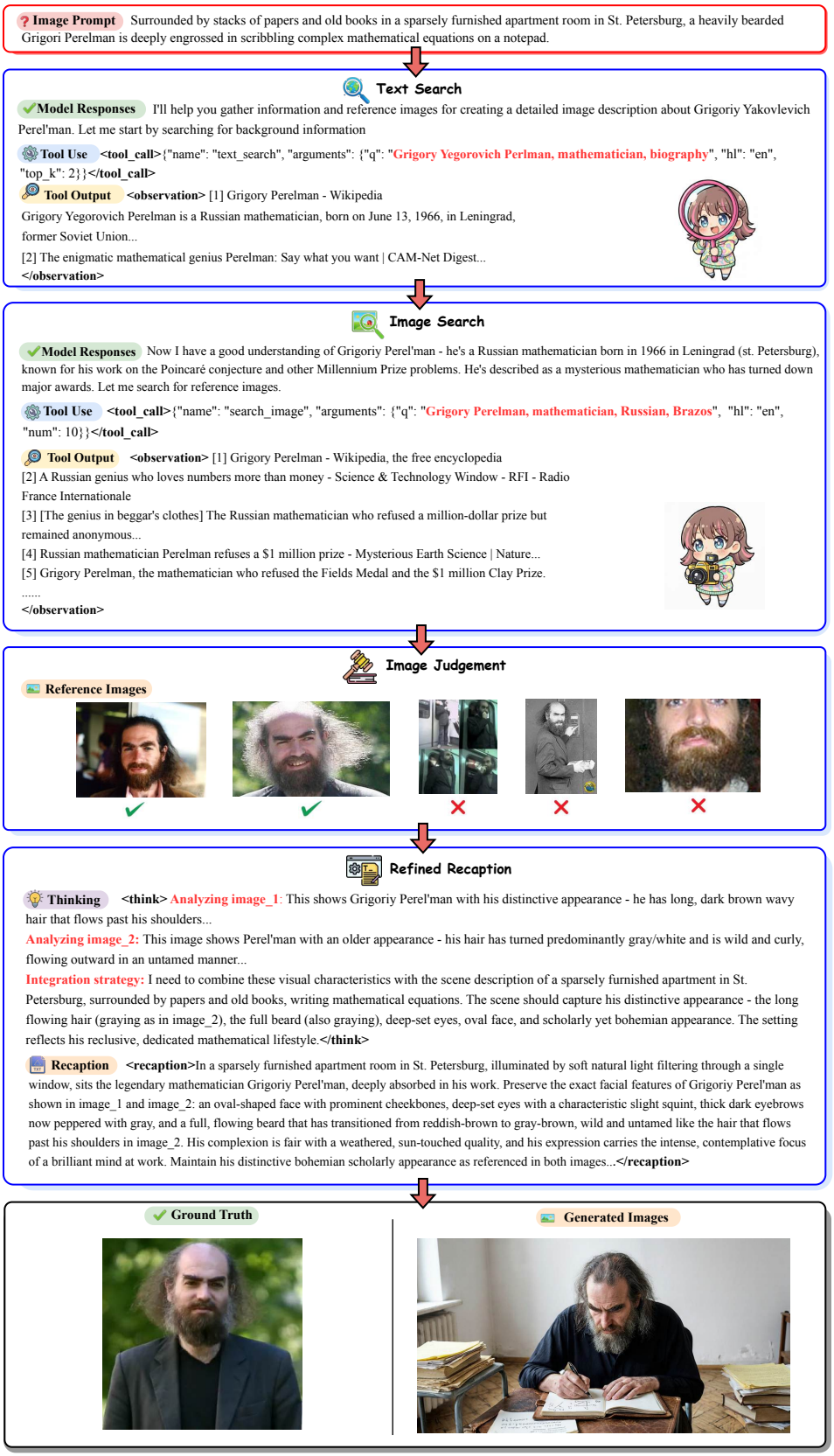


Figure 11: Image generated for Grigory Perelman scribbling mathematical equations.

H FactIP Benchmark Evaluation Showcases

Case 1 Figure 13 presents a representative Seed 2.0 evaluation example for an image generated from a prompt about Gregg Popovich. By comparing the synthesized result against the reference images (GT1 and GT2), the MLLM-based evaluator produces a fine-grained scoring profile together with an interpretable reasoning trace. In particular, the evaluator verifies both the facial resemblance to Popovich and the semantic consistency of basketball-related attributes, such as the Spurs jerseys and the tactical whiteboard. Based on this multi-aspect assessment, the sample receives a high overall score of 9, illustrating the benchmark's ability to jointly measure identity fidelity and contextual relevance.

Case 2 As shown in Fig. 14, this example illustrates how Seed 2.0 handles cases where visual quality remains strong but semantic correctness breaks down. For an image intended to depict Max Weber, the evaluator assigns only a moderate overall score of 5. Although the image is rated highly in terms of clarity (9) and aesthetic appeal (8), the reasoning process identifies two decisive issues: the generated face does not faithfully match the reference appearance—most notably due to the beard discrepancy and the addition of glasses absent from GT1/GT2—and the scene includes a modern microphone that introduces a clear historical anachronism. This case highlights that the benchmark does not merely reward surface-level image quality, but can also capture subtle factual and identity-level inconsistencies.

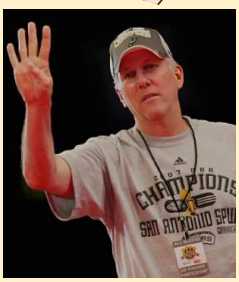
Case 3 Figure 15 provides an example of successful IP-oriented generation under a stylized artistic setting. Given a prompt requesting Habatan rendered in a Baroque oil painting style, the Seed 2.0 evaluator assigns an overall score of 9, reflecting both strong character fidelity and convincing stylistic execution. The accompanying reasoning demonstrates that the evaluator can recognize sophisticated visual properties associated with the Baroque tradition, including dramatic chiaroscuro and dense painterly brushwork, while still preserving the defining attributes of the mascot. Importantly, the analysis also uncovers a subtle but meaningful deviation: the generated image depicts two blue shoes, whereas the ground-truth reference (GT1) specifies one red shoe and one blue shoe. This example underscores the benchmark's capacity for rigorous fine-grained auditing even in highly artistic generations.

Case 4 The failure mode illustrated in Fig. 16 further reveals the strictness of Seed 2.0 in evaluating identity-specific and text-sensitive attributes. For a prompt requiring Scottie Pippen wearing a classic Chicago Bulls jersey, the evaluator assigns a notably low overall score of 2. The reasoning trace points to several critical deficiencies: the generated subject bears little resemblance to Scottie Pippen when compared with GT1 and GT2, the image quality is compromised by substantial visual artifacts, and the jersey details—including the team branding and the iconic number 33—are rendered incorrectly and are largely illegible. This case demonstrates that the benchmark can effectively penalize failures in character identity, visual integrity, and fine-grained textual faithfulness.

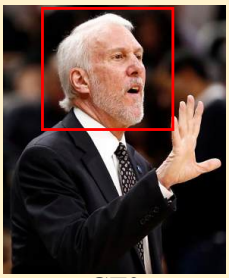
Prompt

Huddled on the sideline of a basketball court during a timeout, Gregg Popovich is intensely drawing a play on a whiteboard, instructing San Antonio Spurs players surrounding him.

Ground Truth

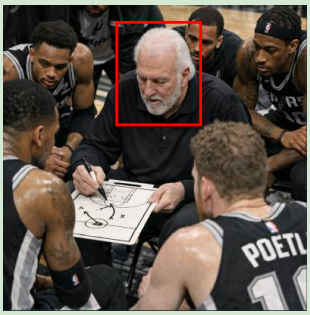


GT1



GT2

Generated Image



MLLM Evaluation



Seed 2.0:

Reasoning Got it, let's evaluate each part step by step. First, clarity: the AS image is sharp, no blur, right? So 9? Wait wait, let's check. Content quality: the prompt says Popovich on sideline timeout drawing play for Spurs players ...

Overall high quality with minimal flaws. Wait let's make sure scores are integers 0-10. Let's adjust if needed? Wait maybe clarity 9, content quality 9, aesthetics 8, text_relevance_ip 10, overall 9.

Rationale The generated image **has high clarity with sharp details of characters and the tactical board**. **Its content fully matches the prompt**: Popovich's likeness is consistent with GT2, it correctly presents the scenario of him drawing a play for surrounding Spurs players during a sideline timeout, all core required elements are included. **The composition is natural and conforms to the aesthetic characteristics of real sports event photos. All text and IP related elements are reasonable and relevant**, the Spurs jersey content is correct, there is no garbled or irrelevant text. The overall quality is high, with only minor non-critical imperfections.

Final Score **Overall: 9**

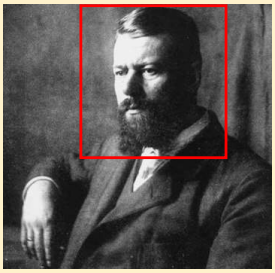
- **clarity: 9 - high clarity with sharp details**
- **content quality: 9 - content fully matches the prompt**
- **aesthetics: 8 - conforms to the aesthetic characteristics of real sports event photos**
- **text relevance: 10 - all text and IP related elements are reasonable and relevant**

Figure 13: MLLM evaluation for Popovich drawing a play.

Prompt

Surrounded by attentive students in a turn-of-the-century German university lecture hall, **Max Weber** is delivering a passionate academic speech from behind a wooden podium.

Ground Truth



GT1



GT2

Generated Image



MLLM Evaluation



Seed 2.0:

Reasoning Got it, let's evaluate each part step by step. First, clarity: The AS image is clear, no blurriness, right? Wait, 9? Wait no, let's see. Wait it's sharp, details are good. Maybe 9? Wait no, wait let's check ... Overall score: 5, while the image is visually well-made and fits the general scene, the inaccurate portrayal of the key figure Max Weber and anachronistic detail greatly reduce its adherence to requirements. Wait wait let's adjust scores properly. Let's make sure all are integers 0-10. Wait let's confirm: clarity: 9, correct, it's very clear. content_quality: 4, as above, correct elements are setting, students, speech action, wrong are subject likeness, anachronistic mic. Aesthetics: 8, yes, looks great, warm lighting, good composition. Text_relevance_ip: 3? Wait or 4? Let's say 3, because the main subject is wrong, which is a core part of the prompt.Overall_score: 5, that makes sense.

Rationale The image has **high sharpness and clear details**, so clarity scores high. It has **good visual performance with appropriate lighting, reasonable composition and historical atmosphere**, so aesthetics scores well. But it has obvious defects: the depiction of Max Weber is **inconsistent with his real appearance** in GT1 and GT2 (with extra glasses, mismatched facial and beard features), and there is an anachronistic modern microphone that did not exist in the turn-of-the-century period. **It deviates greatly from the core requirements of the prompt**, so content quality and text relevance IP scores are low, leading to a medium overall score.

Final Score **Overall: 5**

- ✔ **clarity: 9 - high sharpness and clear details**
- ✘ **content quality: 4 -deviates greatly from the core requirements of the prompt**
- ✔ **aesthetics: 8 - good visual performance**
- ✘ **text relevance: 3 - inconsistent with his real appearance**

Figure 14: MLLM evaluation for Max Weber delivering a passionate speech.

Prompt

A dramatic classical Western oil painting in the Baroque style, illustrating the mascot character Habatan. The round, yellow anthropomorphic phoenix is emerging triumphantly from the fiery ashes and rubble of a destroyed landscape, symbolizing recovery from the Great Hanshin Earthquake. Rich textures, chiaroscuro lighting, grand scale, heavy brushwork.

Ground Truth



GT1



GT2

Generated Image



MLLM Evaluation



Seed 2.0:

Reasoning Got it, let's go through each key step by step. First, clarity: The AS is a clear oil painting, no blur, details are sharp, so that's 10? Wait yeah, looks really clear. Content quality: It's well rendered, the character is consistent with GT1, the background matches the prompt. Wait 9? ... Wait now assign scores properly:\nclarity: 10, right, no blurriness at all, super clear. content_quality: 9, only minor shoe color issue, everything else is great.\naesthetics: 9, really good baroque style, dramatic, beautiful lighting, great texture.\ntext_relevance_ip: 9, IP is almost perfect, only shoe color mismatch, prompt requirements all met (baroque, emerging from rubble, fire, etc.)\noverall_score: 9, since it's almost perfect, only tiny flaw.\nWait wait let's make sure. Oh right GT1 has left blue, right red? Wait GT1: left foot blue, right foot red, AS has both blue, that's the only flaw. Yeah.\nNow structure the JSON correctly, all keys present, scores integers 0-10, rationale clear.\nWait let's make the rationale precise:

Rationale The generated image has **perfect clarity with all details clearly visible**. Its **content quality is high, with only a minor flaw of inconsistent shoe color** (GT1 has one red and one blue shoe while the generated image has two blue shoes). Aesthetically it **fully matches the requested Baroque classical oil painting style**, with heavy brushwork, rich textures, chiaroscuro lighting and dramatic grand scale. It has **high text relevance and IP consistency**: the Habatan mascot aligns with core features of GT1/GT2 (round yellow anthropomorphic phoenix, orange wing patterns, smiling face with pink cheeks), and the scene of the character standing triumphantly on fiery destroyed rubble correctly corresponds to the prompt's symbolic meaning of Great Hanshin Earthquake recovery. The overall performance is excellent with only a tiny imperfection.

Final Score **Overall: 9**

- ✔ **clarity: 10 -** perfect clarity with all details clearly visible
- ✘ **content quality: 9 -** content quality is high
- ✔ **aesthetics: 9 -** fully matches the requested Baroque classical oil painting style
- ✘ **text relevance: 9 -** high text relevance and IP consistency

Figure 15: MLLM evaluation for the mascot character Habatan in the Baroque style.

Prompt

Wearing the iconic red number 33 Chicago Bulls jersey, **Scottie Pippen** is aggressively driving towards the basket during a packed NBA playoff game in the 1990s.

Ground Truth



GT1



GT2

Generated Image



MLLM Evaluation



Seed 2.0:

Reasoning Got it, let's go through each key one by one. First, clarity: the AS image is blurry, face is off, details are messed up, so like 3? Wait no, wait. Wait text_relevance_ip: the jersey text is wrong, not Bulls, number is not 33, it's gibberish. Also the person doesn't look like Pippen. Let's break down each: ... Overall score: The image fails to correctly depict the subject, required jersey details, and has poor visual quality, so overall score 2. Wait wait, but let's make all scores integers, right. Wait wait let me adjust to make it right: Wait clarity: 3, correct, because you can tell it's a basketball player near a hoop but details are blurry and artifacted. Content quality: 2, actually, because the person is not Pippen, which is a core content requirement. Oh right, the main subject is wrong. So content quality 2. Aesthetics: 3? Or 2? Let's say 2. Text_relevance_ip: 0, 100% wrong text and number, no correct branding. Overall score: 2, that fits.

Rationale The AS image has **heavy artifacts and blurry details**, leading to low clarity. While it generally depicts a basketball player in a red jersey near a basket in a packed arena, **the subject does not resemble Scottie Pippen at all**, resulting in poor content quality. **Distorted facial features and garbled visual elements** make it visually unappealing for low aesthetics. **The text on the jersey is completely unrecognizable**, with no correct 'BULLS' branding or number '33', so text relevance IP scores 0. The image fails to meet most core requirements of the prompt, leading to a low overall score.

Final Score **Overall: 2**

- ✔ **clarity: 3 -** **heavy artifacts and blurry details**
- ✘ **content quality: 2 -** **the subject does not resemble Scottie Pippen at all**
- ✔ **aesthetics: 2 -** **distorted facial features and garbled visual elements**
- ✘ **text relevance: 0 -** **the text on the jersey is completely unrecognizable**

Figure 16: MLLM evaluation for Scottie Pippen during an NBA playoff game.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the four main contributions: the Unify-Agent agentic paradigm, architectural insights on UMMs, the FactIP benchmark, and superior performance results. All claims are supported by experimental results in Section 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Sec. 7, including the restricted context length of the base model (Bagel), the one-pass (non-iterative) agentic workflow, and the scope of current evaluation being focused on IP/concept-centric tasks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include formal theoretical proofs. The probabilistic formulation in Section 3.3 (Eq. 3) is a modeling framework rather than a theorem requiring formal proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Training hyperparameters are fully reported in Appendix B and Table 6 (optimizer, learning rate, hardware, steps, etc.). The data pipeline is described in detail in Section 4. Evaluation protocols for all benchmarks are specified in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code, model weights, and the FactIP benchmark will be publicly released. The dataset construction pipeline and all prompts used in data construction are fully described in Section 4 and Appendix F.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details including hardware (64× NVIDIA H20 GPUs), optimizer (AdamW), learning rate (5×10^{-5}), training steps (10K), and all hyperparameters are provided in Appendix B and Table 6. Evaluation settings follow official benchmark protocols as described in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Error bars and statistical significance tests are not reported. Given the large scale of evaluation (e.g., FactIP-Mini with 500 samples, multiple baselines) and the reliance on MLLM-based evaluation which itself incurs significant cost, running multiple independent trials for all experiments would be prohibitively expensive.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B reports that SFT was conducted on 64 NVIDIA H20 GPUs over approximately 10 days for 10K gradient steps using FSDP (HYBRID_SHARD) parallelism. This information is also summarized in Table 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The paper involves no human subjects experiments. Human annotators involved in data curation (Section 4.1.1) were employed in accordance with standard labor practices.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impacts include improved factual image generation for cultural, educational, and creative use. Potential negative impacts include misuse for generating realistic deepfakes or disinformation involving real celebrities and public figures, which is an inherent risk of any advanced image generation system. These are discussed alongside mitigation strategies in Appendix I.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper discusses potential misuse risks of the released model and dataset, including realistic generation of real people (celebrities) and use of web-scraped images. Safeguards including content filtering during data construction, model release restrictions, and benchmark usage guidelines are described in Appendix I.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing models and datasets used (e.g., Bagel, BabelNet, Wikipedia, FLUX, Stable Diffusion) are properly cited with their original references. External APIs (GPT-4o, Gemini, Claude Opus 4.6) are used in accordance with their respective terms of service. All benchmark baselines are cited with their original papers.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Two new assets are introduced: the Unify-Agent model and the FactIP benchmark. The benchmark construction pipeline, taxonomy, filtering criteria, and evaluation protocol are described in detail in Section 4.2 and Appendix E. The model architecture and training procedure are documented in Sections 3, 5, and Appendix B. The assets will be released publicly, together with documentation and usage instructions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

Justification: Human annotators were used for IP collection, seed image curation, and data quality filtering as described in Section 4.1.1. The annotation tasks are described in the paper. Annotators were compensated in accordance with standard practices in the country of data collection.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The human annotator involvement consisted of standard data collection and labeling tasks that do not constitute human subjects research requiring IRB review. No sensitive personal data was collected from participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are a core component of the proposed method. Claude Opus 4.6 is used as the teacher agent for trajectory construction (Section 4.1.2); GPT-4o is used for structured metadata summarization (Section 4.1.1) and as an automatic evaluator on WiSE, KiTTEN, and T2I-FactualBench (Appendix D); Gemini 3 Pro is used for prompt generation; Gemini 3 Flash for visual candidate scoring. All usages are described in detail in Sections 4 and Appendix D/F.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.