MedIGen: Towards Reliable Medical Illustration Generation via Interleaved Introspective Reasoning

Rongsheng Wang[1], **Hongru Zhou**[2], **Ruizhe Zhou**[3], **Haoming Chen**[3], **Zhenyang Cai**[1], **Junying Chen**[1], **Minghao Wu**[1], **Yaofei Duan**[3], **Ziyi Zeng**[1], **Benyou Wang**[1,4,5,6*]

[1]The Chinese University of Hong Kong, Shenzhen [2]Peking Union Medical College Hospital [3]Macao Polytechnic University [4]Shenzhen Loop Area Institute [5]Shenzhen Research Institute of Big Data [6]National Health Data Institute, Shenzhen

Abstract

Medical illustration generation is not merely domain-specific text-to-image (T2I) generation, but a high-constraint visual reasoning problem where visually plausible images can be invalid due to anatomical and structural errors. Existing T2I models largely rely on one-pass generation and lack an internal mechanism to inspect and correct such biomedical inconsistencies. We introduce **MedIGen**, a unified medical illustration generation model that transforms one-pass synthesis into an interleaved introspection-aware refinement process, coupling generation with self-reflection and re-generation through "generate–reflect–refine" cycles. MedIGen is trained through three progressive stages. First, we perform large-scale medical illustration pretraining on 1.21M samples. Second, we conduct mixture training over generation, reflection, and refinement sub-skills. Finally, we apply reinforcement learning with dual-level rewards that optimize both reasoning validity and final visual correctness. We further present **IlluGenBench**, an expert-aligned rubric-driven benchmark with 296 diverse tasks and 9,015 criteria, evaluating scientific accuracy, structural correctness, and semantic alignment beyond coarse visual plausibility. Experiments show that MedIGen substantially outperforms strong open-source unified and reasoning-based baselines, improving the medical reliability of generated illustrations and establishing a new open-source frontier. Data, code, and models are available at `https://github.com/FreedomIntelligence/MedIGen`.

1 Introduction

High-quality visual representations are fundamental to modern medicine. Medical illustrations play an important role in supporting education [29, 17], surgical planning [40], and scientific communication [54]. Unlike natural images, however, their quality is not determined by visual realism alone. A polished illustration can be medically invalid if a surgical step violates procedural logic. Medical illustration generation therefore requires visual generation under strict biomedical constraints, including biomedical semantics, spatial relationships, and structural consistency.

Recent advances in text-to-image generation have substantially reshaped visual synthesis. Diffusion-based methods, exemplified by Stable Diffusion [51] and Qwen-Image [62], have demonstrated remarkable fidelity. Furthermore, the convergence of generative models into unified multimodal architectures [13, 9, 65, 19] has increasingly blurred the boundary between language understanding and visual generation. These advances point to promising instruction-following visual synthesis, but do not ensure reliability under the strict biomedical constraints required for medical illustration.

To address this challenge, we present **MedIGen**, a unified model for medical illustration generation that incorporates interleaved self-reflection and iterative refinement. MedIGen is trained through three progressive stages. First, we perform domain-specific pretraining on 1.21M medical illustration–caption pairs to align biomedical language with fine-grained anatomical visual structures. Second, we

*Correspondence to: Benyou Wang <wangbenyou@cuhk.edu.cn>

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



Figure 1: Medical illustration cases generated by different T2I models. Although these models produce outputs that appear visually plausible, they contain medically meaningful errors. Red block represents the commercial T2I model. Green block represents the open-source T2I model.

conduct supervised mixture training on 150K high-quality samples, decomposing introspection-aware generation into three sub-skills: initial generation, self-reflection, and re-generation. This training enables the model to inspect intermediate outputs and iteratively revise images based on its own textual feedback. Third, we apply GRPO with dual-level rewards that separately supervise the validity of the self-reflection process and the correctness of the final generated image. This objective encourages the model to use reflection as an actionable correction signal rather than as a purely descriptive rationale.

We further introduce **IlluGenBench**, an expert-aligned benchmark for medical illustration generation. IlluGenBench contains 296 tasks spanning diverse medical scenarios, each evaluated with fine-grained rubric criteria designed by domain experts. Unlike coarse text–image alignment metrics, these rubrics assess scientific accuracy, structural correctness, and semantic alignment at the criterion level. Experiments show that MedIGen substantially improves over strong open-source unified and reasoning-based baselines, and narrows the gap to leading proprietary models.

Contributions The main contributions of this work are threefold: **(1)** We introduce **MedIGen**, a unified model that replaces one-pass synthesis with an interleaved "generate–reflect–refine" process for reliable medical illustration generation. **(2)** We propose a progressive training recipe that combines large-scale medical illustration pretraining, mixture training over generation/reflection/refinement skills, and dual-level RL rewards for both reasoning validity and final visual correctness. **(3)** We introduce **IlluGenBench**, an expert-aligned rubric-driven benchmark for criterion-level medical evaluation, and show that MedIGen substantially advances open-source medical illustration generation while narrowing the gap to leading proprietary systems.

2 Motivation

Reliability remains a central challenge for medical illustration generation. Current T2I models often optimize visual plausibility rather than biomedical validity [6, 61], producing polished images that may still violate anatomical, structural, or procedural constraints. In medical contexts, such errors are not cosmetic: they can misrepresent biological knowledge and undermine educational or clinical communication. This limitation is amplified by one-pass generation pipelines, which lack an explicit mechanism to verify and correct medically meaningful errors.

Case Study We first conduct a qualitative study on five representative medical illustration tasks. As shown in Figure 1, models often generate images that appear polished at first glance, yet violate medically meaningful constraints such as anatomical placement, structural connectivity, or procedural logic. Open-source models exhibit more frequent and severe errors, but even leading commercial systems fail to consistently satisfy expert expectations.

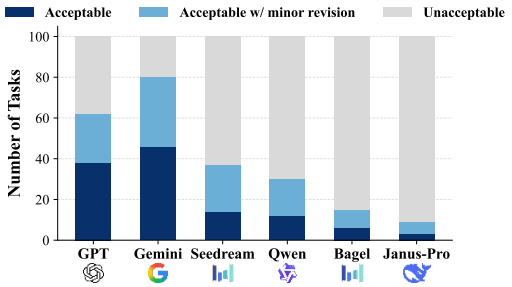


Figure 2: Quantitative evaluation of medical illustration generation across different models on 100 tasks. Outputs are categorized by biomedical experts as *acceptable*, *acceptable with minor revisions*, or *unacceptable*.

Quantitative Experiment We further scale the analysis to 100 medical illustration tasks and collect judgments from three experts. Figure 2 shows that commercial models consistently outperform open-source baselines, but none achieves a fully correct rate above 50%. Many outputs are unacceptable, while others require only targeted anatomical or structural revisions.

Key Findings The results reveal two insights. **(1)** Visual plausibility does not ensure biomedical fidelity, as even leading commercial models produce anatomical, structural, and procedural errors. **(2)** These errors are often localized rather than global, concentrating on specific anatomical, structural, and procedural details.

3 MedIGen: Reliable Medical Illustration Generation Model

3.1 Preliminaries and Problem Formulation

Unified Autoregressive Architecture We construct our framework upon *Janus-Pro* [9], a unified multimodal model. Janus-Pro adopts an autoregressive architecture that discretizes both text and images into unified token sequences. This formulation allows a single Transformer backbone to seamlessly handle both visual understanding and generation via next-token prediction. Crucially, this bidirectional capability is the cornerstone of our method: it empowers the model with introspective potential within a shared parameter space.

Interleaved Introspective Reasoning Let T_{in} denote the input text prompt. We formulate illustration generation as a progressive interleaved introspection-aware refinement process, where the generated image is iteratively improved through intermediate evaluation and feedback.

At iteration k, the generated image is denoted as $I^{(k)}$. The process starts by synthesizing an initial image using a text-to-image generator, $I^{(0)} = \mathcal{G}(T_{\text{in}})$. After each generation step, the model performs a self-reflection that evaluates the semantic consistency and anatomical correctness between T_{in} and the current image $I^{(k-1)}$, producing a brief textual feedback $T_{\text{ref}}^{(k)} = \mathcal{R}(T_{\text{in}}, I^{(k-1)})$.

Conditioned on this feedback, the generator is re-invoked with the original prompt, previous image, and reflection output, yielding an updated image $I^{(k)} = \mathcal{G}(T_{\text{in}}, I^{(k-1)}, T_{\text{ref}}^{(k)})$. Each iteration improves semantic alignment and scientific validity. The process terminates once the image satisfies the input specification or a maximum iteration budget (default is 3) is reached.

3.2 Progressive Training

3.2.1 Stage 1: Domain-Aligned Visual–Semantic Pretraining

To bridge the domain gap between Janus-Pro-7B's [9] natural image pre-training and specialized medical imagery, we conduct visual-semantic alignment using the BIOMEDICA dataset [37]. Specifically, we curate a subset of **1.21 million medical illustration-caption pairs** from the original 24 million corpus. Training on this domain-specific subset efficiently adapts the visual representations, ensuring perception of anatomical features prior to reasoning training. The details of data construction are provided in the Appendix C.1. The data ratio between understanding and generation is set to 3:7.

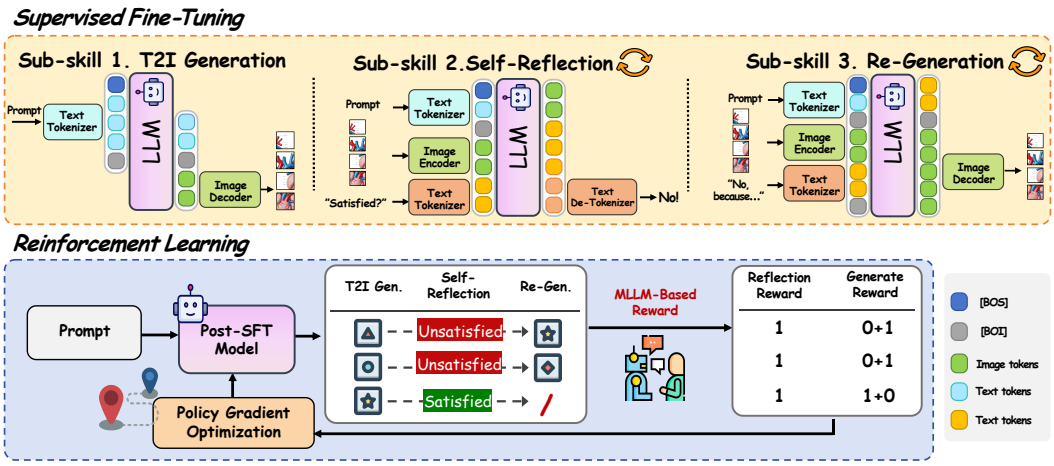


Figure 3: **Overview of MedIGen training.** Following domain-aligned pretraining, MedIGen is optimized through supervised mixture fine-tuning over generation, reflection, and refinement skills, and further improved with dual-level reinforcement learning that rewards both introspective validity and final visual correctness.

3.2.2 Stage 2: Unlocking Introspective Reasoning via Mixture Training

We decompose introspection-aware refinement into three sub-skills: text-to-image generation, self-reflection, and re-generation.

Text-to-Image Generation To improve fine-grained anatomical modeling, we optimize the unified autoregressive generation objective on a high-quality medical illustration dataset with strict anatomical constraints. Given prompt $\mathbf{x}$ and image tokens $\mathbf{z} = (z_1, \ldots, z_T)$, training minimizes $\mathcal{L}_{\text{gen}} = -\mathbb{E}_{(\mathbf{x},\mathbf{z})} \sum_{t=1}^{T} \log p_\theta(z_t \mid z_{<t}, \mathbf{x})$.

Self-Reflection Given instruction $\mathbf{x}$ and generated image $\mathbf{z}$, the model learns to assess semantic consistency and anatomical correctness, producing feedback that identifies medically meaningful errors.

Re-Generation Conditioned on the original instruction, accumulated reflections $\mathbf{r}_{<k}$, and the previous image, the model generates refined image tokens $P(\mathbf{z}_k \mid \mathbf{x}, \mathbf{r}_{<k}, \mathbf{z}_{k-1})$, allowing reflective reasoning and visual evidence to guide correction.

We jointly train these sub-skills with a **mixture training** strategy, integrating them into a unified introspection-aware refinement process.

3.2.3 Stage 3: Dual-Level Reinforcement Learning for Reflection-Guided Refinement

To move the model beyond imitation toward more reliable introspection-aware refinement. We adopt Group Relative Policy Optimization (GRPO) [52] with *Dual-Level Rewards* to encourage accurate self-evaluation and its causal impact on illustration refinement.

Dual-Level Rewards We design a dual reward for each trajectory i. Each trajectory consists of an initial generated image $\mathcal{I}_i^{(0)}$, a self-evaluation $s_i \in \{0, 1\}$, and a final image $\mathcal{I}_i^{(K)}$ obtained after multiple rounds of self-reflection and re-generation, where K denotes the iteration index at which the trajectory terminates. We use *Qwen3-VL-30B-A3B* as the external evaluator.

- *Generation Rewards*: An external vision–language model evaluates the alignment between each generated image and the input prompt $\mathbf{x}$, yielding scores $R(\mathcal{I}_i^{(0)}, \mathbf{x})$ and $R(\mathcal{I}_i^{(K)}, \mathbf{x})$ for the initial and final images, respectively.
- *Understanding Reward*: To ensure that the model's self-evaluation reflects true generation quality, we define a consistency-based understanding reward as $R_i^{\text{und}} = 1 - \left| R(\mathcal{I}_i^{(0)}) - s_i \right|$. This term encourages the model to produce a self-check output s_i that is consistent with the external evaluator.

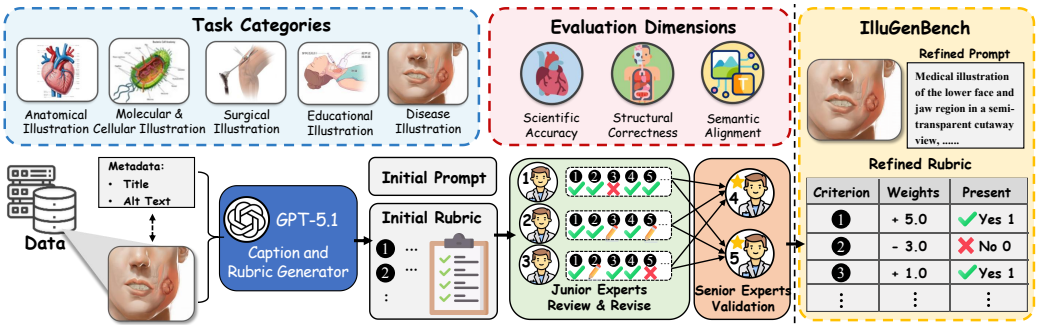


Figure 4: **Benchmark Construction.** (a) **Initialization**: GPT-5.1 utilizes metadata and illustration across five medical categories to synthesize initial prompts and rubrics. (b) **Hierarchical Expert Refinement**: Review process by junior and senior experts validates fidelity, yielding the final refined prompts and weighted scoring criteria.

Policy Gradient Optimization For each prompt, we sample G trajectories and compute a normalized advantage for each stage $k \in \{\text{init}, \text{und}, \text{final}\}$:

$$A_i^{(k)} = \frac{R_i^{(k)} - \mu(\{R_j^{(k)}\}_{j=1}^G)}{\sigma(\{R_j^{(k)}\}_{j=1}^G) + \epsilon}.$$

Each stage is optimized only with its own advantage, yielding a clipped GRPO objective averaged over stages:

$$\mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|k|}\sum_{t\in k}\min\Big(\rho_{i,t}A_i^{(k)}, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_i^{(k)}\Big) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right].$$

This decouples credit assignment for generation, reflection, and refinement.

4 IlluGenBench: Rubric-driven Evaluation for Medical Illustration

IlluGenBench consists of **296 medical illustration generation tasks** spanning five categories. Each task is paired with an average of approximately 30 task-specific rubric criteria. All tasks and rubric criteria are developed by multiple experts with the assistance of GPT-5.1 and iteratively refined to ensure clarity, medical relevance, and expert consistency. The benchmark is designed to reflect real-world medical illustration generation scenarios. In total, IlluGenBench contains **9,015 unique rubric criteria**, enabling fine-grained evaluation of medical illustration generation across three dimensions. The construction pipeline of IlluGenBench is shown in Figure 4.

4.1 Task Categories and Evaluation Dimensions

Task Categories The benchmark is organized into five task categories including **(1)** Anatomical Illustration; **(2)** Surgical Illustration; **(3)** Molecular and Cellular Illustration; **(4** Educational and Patient Communication Illustration; and **(5)** Disease and Disorder Illustration, covering core medical illustration scenarios.

Evaluation Dimensions We evaluate medical illustration generation along three senior expert-defined dimensions that reflect medical correctness. Specifically, we consider **(1) Scientific Accuracy**, which assesses consistency with established medical knowledge; **(2) Structural Correctness**, which evaluates the integrity and spatial relationships of anatomical and biological components; and **(3) Semantic Alignment**, which measures how faithfully the illustration follows the given medical instruction.

4.2 Benchmark Construction

Data Selection We construct the benchmark using the dataset described in Section C.3. To ensure diversity and balance, we first use GPT-5.1 to categorize dataset according to the predefined task

Table 1: **Performance comparison of different T2I generation models on IlluGenBench. Bold** indicates the best performance. Commercial models are shown in gray for reference only. Underlined indicate the second-best performance. Commercial models are shown in gray for reference only due to undisclosed details. *B* indicates billion, and *A* indicates activated parameters.

Model	Parameters	Scientific Accuracy↑	Structural Correctness↑	Semantic Alignment↑	Average↑
Commercial T2I Generation Models (Reference Only)					
GPT-Image-1	×	0.843	0.812	0.847	0.835
GPT-Image-1.5	×	0.849	0.811	0.852	0.838
GPT-Image-2	×	0.800	0.761	0.768	0.777
Gemini-2.5-Flash-Image	×	0.733	0.676	0.789	0.734
Gemini-3-Pro-Image	×	0.879	0.849	0.890	0.873
Gemini-3.1-Flash-Image	×	0.869	0.852	0.886	0.870
Seedream-4.5	×	0.787	0.692	0.825	0.769
Kling-Image-v2.1	×	0.173	0.129	0.272	0.190
Open-Source T2I Generation Models					
SDXL [45]	3.5B	0.103	0.061	0.170	0.111
Playground-v2.5 [31]	3.5B	0.063	0.043	0.147	0.083
FLUX.1-dev [30]	12B	0.375	0.324	0.476	0.391
Stable-Diffusion-3.5 [55]	8.1B	0.220	0.152	0.267	0.213
Chroma1-HD [50]	8.9B	0.417	0.332	0.506	0.419
HiDream-I1-Full [3]	17B	0.247	0.212	0.311	0.256
Lumina-Image-2.0 [47]	2.6B	0.308	0.239	0.404	0.317
Qwen-Image [61]	20B	0.434	0.344	0.517	0.432
Qwen-Image-2512 [61]	20B	**0.644**	**0.565**	<u>0.590</u>	<u>0.601</u>
Unified Understanding and Generation Models					
Janus-Pro-1B [9]	1B	0.174	0.110	0.370	0.217
Janus-Pro-7B [9]	7B	0.298	0.224	0.463	0.328
Janus-4o [6]	7B	0.416	0.318	0.566	0.433
BAGEL [13]	14B (A7B)	0.350	0.301	0.521	0.390
BLIP3o-NEXT [5]	3B	0.319	0.266	0.445	0.343
UniWorld-V1 [34]	19B	0.265	0.202	0.416	0.294
Emu3.5 [11]	8B	0.306	0.257	0.470	0.344
Show-o2 [65]	7B	0.244	0.203	0.435	0.273
GLM-Image [68]	16B	0.492	0.430	0.552	0.491
T2I Reasoning Models					
GoT [16]	6B	0.287	0.196	0.319	0.262
Janus-Pro-R1 [43]	7B	0.014	0.008	0.135	0.052
Uni-CoT (v0.2) [46]	14B (A7B)	0.384	0.321	0.506	0.413
T2I-R1 [28]	7B	0.258	0.186	0.424	0.289
Our Models					
MedIGen	7B	<u>0.606</u>	<u>0.537</u>	**0.711**	**0.618**

categories. Three junior experts then independently sample illustrations, prioritizing copyright-free professional sources and using web data only when necessary. The selected illustrations are merged and subsequently reviewed and filtered by two senior experts to form the final benchmark set. The selected illustrations are removed from the training set.

Caption Generation We provide each illustration together with its associated metadata to GPT-5.1 and prompt the model to generate a descriptive caption that faithfully summarizes the medical illustration. The prompts can be found in the Appendix F.1.

Rubric Criteria Creation To construct evaluation rubrics, we provide each illustration together with its generated descriptive caption to GPT-5.1 and prompt it to produce a set of customized criteria that explicitly define the requirements of a qualified medical illustration. Each rubric criterion is assigned a numerical weight in the range $[-5, 5]$ to reflect its relative importance. Positive weights reward desirable attributes, whereas negative weights penalize common failure modes such as anatomical inaccuracies or structural inconsistencies. The prompts can be found in the Appendix F.2.

Expert Check and Refinement Multiple experts further review and refine the outputs from the previous stages. Specifically, three junior experts first examine and revise the generated descriptive

captions and their corresponding rubric criteria to ensure medical correctness, clarity, and internal consistency. For conflicting or ambiguous criteria, consensus is reached through discussion. The revised results are then reviewed by two senior experts, who perform a final round of validation and refinement to resolve remaining issues and ensure expert-level consistency and reliability.

4.3 Evaluation Methodology

Under the MLLM-as-a-judge [4] setting, each generated output is evaluated against all associated criteria and reference images (human-created illustrations). For each criterion, the judging model produces a binary decision indicating whether it is satisfied. Both positive and negative criteria are considered, with negative weights applied only when the corresponding failure conditions are triggered.

The task score is computed as a normalized weighted sum:

$$S_k = \frac{\sum_{r_i \in C} w_{r_i} \, \mathbf{1}(r_i \text{ is satisfied})}{\sum_{r_i \in C,\, w_{r_i} > 0} w_{r_i}}, \tag{1}$$

where C denotes the set of all rubric criteria, w_{r_i} is the (possibly negative) weight assigned to criterion r_i, and $\mathbb{I}[\cdot]$ is an indicator function that equals 1 if the judging model determines criterion r_i to be satisfied, and 0 otherwise. The normalization term corresponds to the sum of positive weights, representing the maximum achievable score.

5 Experiments

5.1 Experimental Setup

We freeze the vision encoder to maintain a stable visual codebook, while fully fine-tuning the LLM and all task-specific modules. Further training details are provided in the Appendix D.2. For automatic evaluation on IlluGenBench, we use *Gemini-3-Pro* as an automated judge. For human evaluation, we randomly select 100 tasks. Model outputs are randomly ordered, and evaluators provide an overall judgment. Results are reported as preference ratios. Detailed evaluation protocols are provided in Appendix D. At inference time, MedIGen controls the computation–quality trade-off through a tunable maximum generation rounds (MGR) hyperparameter. We set the default to MGR$=3$, while MGR$=1$ disables additional re-generation after the initial generation pass. An analysis of this hyperparameter is provided in Appendix E.2.

5.2 Evaluation Results of MedIGen

Automatic Evaluation Results Table 1 reports the main results on IlluGenBench. First, **MedIGen achieves a $+0.29$ absolute improvement in average score over Janus-Pro-7B**. This gain reflects the effect of progressive training. Second, **MedIGen exhibits strong parameter efficiency**. As shown in Table 1 and Figure 8, the 7B model surpasses substantially larger models such as Qwen-Image-2512 after the final RL stage, indicating that introspective reasoning is more effective than naive parameter scaling. Finally, **these results establish a new state of the art among open-source methods**. Overall, MedIGen moves beyond simple text-to-image alignment and achieves high-level scientific accuracy efficiently.

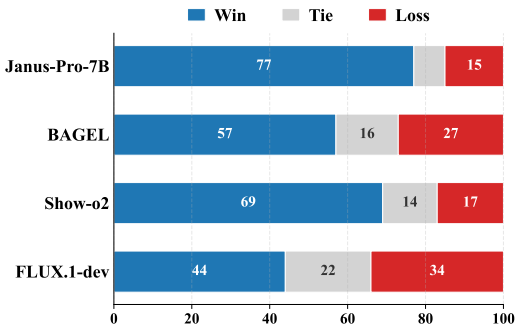


Figure 5: **Human preference evaluation.** Wins indicate cases where evaluators preferred MedIGen over the compared model.

Human Evaluation Results The results in Figure 5 show that MedIGen consistently aligns more closely with human preferences than all baselines, achieving higher win rates with few ties. Its advantages are clearly distinguishable to human evaluators and remain consistent with the trends observed in automatic evaluations.

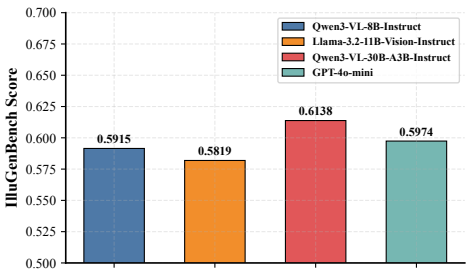


Figure 6: Performance of RL with different Reward Models.

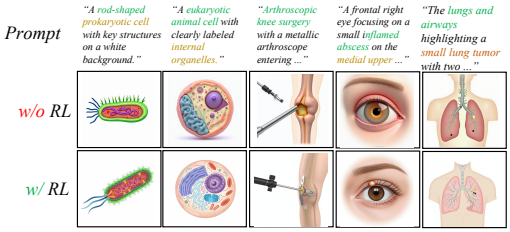


Figure 7: Qualitative comparison of MedIGen before and after RL.

5.3 Ablation Study

Training Stages. Table 2 shows that the three training stages play complementary roles. Stage 1 improves scientific accuracy and structural correctness through medical visual grounding, Stage 2 further enhances these metrics and semantic alignment via introspective error correction, and Stage 3 delivers consistent gains across all metrics by stabilizing reasoning with dual-level RL. Overall, these stages enable reliable medical illustration generation.

Table 2: **Performance of MedIGen over successive training stages across evaluation dimensions.** S1–S3 denote successive training stages built incrementally upon the original model.

	Sci. Acc.	Struct. Corr.	Sem. Align.	Avg.
Original	0.30	0.22	0.46	0.33
+S1	$0.52_{+0.22}$	$0.47_{+0.24}$	$0.64_{+0.18}$	$0.54_{+0.21}$
++S2	$0.57_{+0.05}$	$0.52_{+0.05}$	$0.68_{+0.04}$	$0.59_{+0.05}$
+++S3	$0.61_{+0.04}$	$0.54_{+0.02}$	$0.71_{+0.04}$	$0.62_{+0.03}$

Reward Models Design. As shown in Figure 6, we evaluate different vision-language models as reward models using the IlluGenBench score. Qwen3-VL-30B-A3B-Instruct achieves the best performance, indicating that stronger multimodal understanding leads to more reliable and fine-grained reward signals, particularly for assessing text–image semantic consistency.

RL Training. We analyze the impact of RL on model performance. As shown in Table 2, the SFT primarily provides foundational improvements over the baseline model. However, this stage tends to encourage imitation of the refinement format, often relying on rote memorization with limited transferability. In contrast, RL enables substantially stronger generalization, leading to **marked improvements in image fidelity**, as shown in Figure 7, and empirically validating the effectiveness of the introspection-aware refinement mechanism.

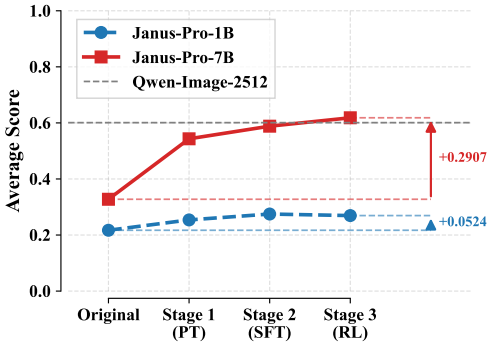


Figure 8: Performance comparison of Janus-Pro-1B and Janus-Pro-7B across progressive training stages.

Model Scale. Figure 8 illustrates the impact of model scale on performance. **Janus-Pro-1B** shows limited gains in early stages and degrades at the final stage, indicating that insufficient model capacity hinders stable learning and effective utilization of advanced training signals. In contrast, **Janus-Pro-7B** achieves substantial and consistent performance gains as training progresses. Overall, these results highlight that scaling model capacity is critical for unlocking the full benefits of multi-stage training and introspection-aware refinement.

Table 5: **Performance of MedIGen across categories and rubric dimensions.** *Sci. Acc.*: Scientific Accuracy; *Struct. Corr.*: Structural Correctness; *Sem. Align.*: Semantic Alignment.

Category	Sci. Acc.	Struct. Corr.	Sem. Align.	Avg.
Educational	0.6590	0.6602	0.7871	0.7021
Anatomical	0.5223	0.5124	0.6328	0.5558
Surgical	0.6055	0.4343	0.6933	0.5777
Disease	0.4326	0.3560	0.5930	0.4605
Molecular	0.8272	0.7936	0.8550	0.8253

Table 3: Macro-F1 agreement between MLLM judges and human experts.

Generation Model	Judge Model	
	GPT-5.1	Gemini-3-Pro
Gemini-3-Pro-Image	0.741	**0.782**
MedIGen	0.735	**0.760**

Table 4: Impact of reference images on judge–expert agreement.

Generation Model	Setting	Macro F1
Gemini-3-Pro-Image	Ref	**0.782**
Gemini-3-Pro-Image	Non-Ref	0.642
MedIGen	Ref	**0.760**
MedIGen	Non-Ref	0.634

Alignment Between Human and LLM Judges. We evaluate the consistency between GPT-5.1 and Gemini-3-Pro as judges and human experts. Table 3 shows that Gemini-3-Pro achieves the highest agreement with human experts, while the remaining gap suggests that there is still room for improvement.

Impact of Reference Images. We evaluate the impact of reference illustrations on MLLM-based evaluation. As shown in Table 4, providing reference images substantially improves agreement between Gemini-3-Pro (judge) and human experts for MedIGen-generated illustrations. Removing reference images (*Non-Ref*) results in a sharp drop in consistency, highlighting the importance of reference images for stable MLLM-based evaluation.

Impact of Expert Refinement. We evaluate the impact of initially MLLM-generated rubrics versus expert-refined rubrics on evaluation alignment. Figures 9 (a) and (b) show that expert-refined rubrics outperform verbose MLLM-generated rubrics, indicating that expert curation cannot be replaced by automated expansion and that clarity stems from precision rather than elaboration.

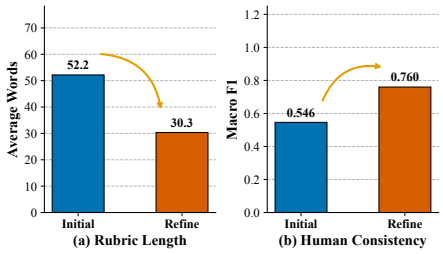


Figure 9: Comparison between expert-refined concise rubrics and verbose MLLM-generated rubrics.

5.4 Analysis of Failure Cases

Our analysis reveals performance variations across medical illustration task categories, as shown in Table 5. While the model demonstrates proficiency in generating structured concepts like Molecular illustrations (*Avg. score: 0.8253*), it encounters typical difficulties with the abstract and pathologically complex Disease category (*Avg. score: 0.4605*).

Moreover, we conduct an expert audit of failures and decompose the pipeline into self-reflection and re-generation stages. The results, detailed in Appendix E.3, show that **the main bottleneck lies in re-generation rather than error detection**: MedIGen often identifies pathological inaccuracies but fails to translate the critique into a corrected visual structure.

5.5 Analysis of Computational Cost and Quality

Table 6 compares the inference cost of MedIGen with its Janus-Pro-7B. MedIGen applies re-generation selectively: each image is first checked by self-reflection, and refinement is triggered only when instruction violations or anatomical errors are detected. In our experiments, 80% of prompts are accepted after the first pass, incurring only modest overhead over Janus-Pro-7B (+1.58s latency

Table 6: **Inference efficiency comparison between Janus-Pro-7B and MedIGen.** Latency, peak memory, and FLOPs are measured per image. "First-pass accepted" denotes MedIGen samples that are accepted by self-reflection without triggering re-generation.

Setting	Latency	Peak Memory	FLOPs	Avg. Gen. Rounds	Fraction
Janus-Pro-7B	12.63 s	15.18 GB	26.46 T	1.00	100%
MedIGen (first-pass accepted)	14.21 s	15.44 GB	26.89 T	1.00	80%
MedIGen (overall average)	21.83 s	16.06 GB	57.91 T	1.41	100%

and +0.43T FLOPs). The higher overall average cost is therefore driven by harder cases that require additional refinement, indicating that **MedIGen allocates extra test-time computation adaptively rather than uniformly**.

6 Related Works

Unified Understanding and Generation Models. Unified multimodal understanding and generation models have attracted significant attention from the research community. Existing approaches can be broadly categorized into three paradigms. (1) Autoregressive models [63, 8, 38, 48, 56, 59], which adopt a next-token prediction paradigm to generate both text and image tokens within a single unified model. (2) Models with additional diffusion modules [15, 64, 44, 58], which typically combine a pre-trained large language model (LLM) backbone with an external diffusion model. In these approaches, the LLM provides semantic conditions that guide the diffusion process for image generation. (3) Unified integrated transformer models [12, 33, 39], which tightly integrate LLMs and diffusion models within a single transformer architecture.

Reasoning in Image Generation. Text-based reasoning models have demonstrated substantial progress in solving a wide range of real-world tasks [27, 20, 7]. More recently, interleaved reasoning has emerged as a promising paradigm that integrates non-text modalities into multi-step reasoning processes to address increasingly complex problems. In the context of text-to-image (T2I) generation, recent studies [16, 46, 26, 22, 43] investigate whether incorporating text-based reasoning or introspective mechanisms can improve the quality and reliability of generated images. These approaches primarily aim to enhance general-purpose image generation by embedding reasoning capabilities into unified multimodal models. A extended related work are provided in Appendix B.

7 Conclusion

Medical illustrations are essential for medical education, patient communication, and scientific visualization, yet anatomical and structural inaccuracies can compromise their reliability. To address this challenge, we introduce **MedIGen**, an introspection-driven unified model that improves the accuracy of generated medical illustrations through interleaved refinement. MedIGen is trained progressively with domain-aligned pretraining, introspective sub-skill mixture learning, and reinforcement learning with dual-level rewards. We further introduce IlluGenBench, the first expert-aligned, rubric-driven benchmark for scientific accuracy in medical illustration. Experiments show that MedIGen consistently outperforms competitive baselines, promoting reliable medical illustration generation.

Acknowledgements and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0203000); the National Natural Science Foundation of China (Grant No. 72495131); the Shenzhen Science and Technology Program—General Program of Basic Research (Grant No. JCYJ20250604141206009); the Shenzhen Medical Research Fund (Grant Nos. B2503005 and C2406002); the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (Grant 2023B1212010001 and ZSZDSYS20251011093302004); the 1+1+1 CUHK–CUHK(SZ)–GDSTC Joint Collaboration Fund (Grant No. 2025A0505000078); and the Research Task Assignment Project from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (No. GML-26420011). The authors declare no competing interests.

References

[1] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

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

[3] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

[4] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025.

[6] Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025.

[7] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

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

[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] Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, et al. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing. *arXiv preprint arXiv:2509.24900*, 2025.

[11] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 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. *arXiv preprint arXiv:2505.14683*, 2025.

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[15] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024.

[16] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

[17] Avi A Gajjar, Shrey V Patel, Raj Swaroop Lavadi, Rida Mitha, Rohit Prem Kumar, Tavis Taylor, Galal A Elsayed, D Kojo Hamilton, and Nitin Agarwal. Art and neurosurgery: the importance of medical illustration. *World Neurosurgery*, 181:82–89, 2024.

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

[19] Google. Introducing gemini 2.5 flash image: Our state-of-the-art image model. `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/`, 2025.

[20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[21] Meng-Hao Guo, Jiajun Xu, Yi Zhang, Jiaxi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong, Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, et al. R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation. *arXiv preprint arXiv:2505.02018*, 2025.

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

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

[26] Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*, 2025.

[27] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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

[29] Alina Krasnoryadtseva, Nicola Dalbeth, and Keith J Petrie. The effect of different styles of medical illustration on information comprehension, the perception of educational material and illness beliefs. *Patient education and counseling*, 103(3):556–562, 2020.

[30] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[31] 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. *arXiv preprint arXiv:2402.17245*, 2024.

[32] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Lingyu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *Advances in Neural Information Processing Systems*, 37:18535–18556, 2024.

[33] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

[34] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[36] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.

[37] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19724–19735, 2025.

[38] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*, pages 26439–26455, 2024.

[39] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.

[40] Kimura Masashi, Kondo Yutaro, Hyodo Mizuki, Kataoka Tatsuya, and Hashimoto Kengo. A simple method for creating medical illustrations using tablets. *Cureus*, 15(6), 2023.

[41] Jordan Meyer, Nick Padgett, Cullen Miller, and Laura Exline. Public domain 12m: A highly aesthetic image-text dataset with novel governance mechanisms. *arXiv preprint arXiv:2410.23144*, 2024.

[42] OpenAI. Introducing our latest image generation model in the api. `https://openai.com/index/image-generation-api/`, 2025.

[43] Kaihang Pan, Yang Wu, Wendong Bu, Kai Shen, Juncheng Li, Yingting Wang, Yunfei Li, Siliang Tang, Jun Xiao, Fei Wu, et al. Unlocking aha moments via reinforcement learning: Advancing collaborative visual comprehension and generation. *arXiv preprint arXiv:2506.01480*, 2025.

[44] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

[45] 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.

[46] Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025.

[47] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025.

[48] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[50] Lodestone Rock. Chroma1-hd. `https://huggingface.co/lodestones/Chroma1-HD`, 2025.

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[52] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[53] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer, 2020.

[54] Yogesh Ashok Sontakke. Medical illustration–fusion of art, technology, and medical education. *CosmoDerma*, 2, 2022.

[55] Stability AI. Introducing stable diffusion 3.5. `https://stability.ai/news/introducing-stable-diffusion-3-5`, October 2024. Accessed: 2026-02-06.

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

[57] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[58] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

[59] 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.

[60] 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.

[61] 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.

[62] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[63] 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.

[64] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first ICML*, 2024.

[65] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

[66] 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.

[67] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

[68] Z.AI. Glm-image: Auto-regressive for dense-knowledge and high-fidelity image generation. `https://z.ai/blog/glm-image`, 2026. Accessed: 2026-02-06.

[69] Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, Xiangyu Zhao, Xinyu Fang, Shengyuan Ding, Ziheng Jia, Jiahao Xiao, Ye Shen, Yushuo Zheng, Xiaorong Zhu, Yalun Wu, Ziheng Jiao, Wei Sun, Zijian Chen, Kaiwei Zhang, Kang Fu, Yuqin Cao, Ming Hu, Yue Zhou, Xuemei Zhou, Juntai Cao, Wei Zhou, Jinyu Cao, Ronghui Li, Donghao Zhou, Yuan Tian, Xiangyang Zhu, Chunyi Li, Haoning Wu, Xiaohong Liu, Junjun He, Yu Zhou, Hui Liu, Lin Zhang, Zesheng Wang, Huiyu Duan, Yingjie Zhou, Xiongkuo Min, Qi Jia, Dongzhan Zhou, Wenlong Zhang, Jiezhang Cao, Xue Yang, Junzhi Yu, Songyang Zhang, Haodong Duan, and Guangtao Zhai. Large multimodal models evaluation: A survey. `https://github.com/aiben-ch/LMM-Evaluation-Survey`, 2025. Project Page: AIBench, available online.

A Limitation and Impact Statements

Limitation MedIGen achieves strong overall performance but still has limitations. Although interleaved introspective reasoning markedly improves anatomical and structural fidelity, the model does not guarantee complete correctness across all medical illustration tasks. Performance degrades in particularly complex or abstract cases, such as disease and pathological illustrations requiring deep causal and relational understanding. These challenges point to opportunities for future improvement through tighter integration of structured biomedical knowledge, rather than shortcomings of the introspective reasoning paradigm itself.

Impact Statements This paper studies reliable medical illustration generation, a high-stakes domain where visual errors can lead to misunderstanding or misuse. We introduce MedIGen, which integrates introspective reasoning into text-to-image generation, and IlluGenBench, an expert-aligned benchmark emphasizing scientific accuracy over visual appeal.

The main positive impact is reducing hallucinations and improving structural correctness in medical illustrations, benefiting medical education, communication, and scientific visualization. At the same time, generated images may still contain errors and should not be used as authoritative clinical references. Our framework is intended as an assistive tool, with explicit reasoning and rubric-based evaluation designed to make potential failures more transparent.

More broadly, this work encourages reasoning-aware generation and domain-specific evaluation for safer deployment of generative models in high-risk applications.

B Extended Related Work

Text-to-Image Generation. Text-to-image (T2I) generation has been a major research focus in the field of generative AI in recent years. Mainstream approaches include techniques based on diffusion models [24, 14, 30] or autoregressive models [49, 57], which excel at generating images with high aesthetic quality and semantic consistency. With the advancement of multimodal large language models (MLLMs) [2] that focus solely on understanding, unified multimodal large models [65, 12, 60, 9] have become an emerging research hotspot. These models are capable of supporting both multimodal understanding and image generation within a single framework. The core advantage of unified multimodal models lies in their powerful multimodal understanding and reasoning capabilities, which significantly enhance image generation. Cutting-edge models such as GPT-Image-1 [42] and Gemini-2.5-Flash-Image [19] demonstrate remarkable abilities in generating natural and visually pleasing images. They are widely used for creating artworks, advertising visuals, and virtual scene generation, covering a diverse range of application scenarios.

Text-to-Image Generation Benchmarks. Early evaluations of text-to-image (T2I) generation primarily relied on image similarity metrics such as FID [23]. However, these methods struggle to capture the high-level semantics of generated images. Current evaluation approaches mainly focus on the literal alignment between text and image, as seen in benchmark tests like GenEval [18] and DPG-Bench [25], as well as metrics such as CLIPScore [23] and VQAScore [36]. Inspired by studies on reasoning capabilities in both pure-text and multimodal tasks [21, 67], recent research has increasingly shifted toward reasoning-based generation evaluation, often employing multimodal large language models (MLLMs) as evaluators [69]. Nevertheless, these methods tend to emphasize world knowledge or commonsense reasoning, and their research scope remains largely limited to natural or synthetic images.

Text-to-Image Generation Datasets. Early image–text paired datasets such as MS-COCO [35] and TextCaps [53] laid the foundation for image generation and understanding tasks. Subsequently, researchers collected larger-scale datasets from the internet (e.g., DenseFusion-1M [32] and Public-Domain-12M [41]), significantly increasing sample size and textual diversity. Meanwhile, leveraging the generative capabilities of GPT-4o, datasets like ShareGPT-4o-Image [6], Echo-4o-Image [66] and OpenGPT-4o-Image [10] achieved notable improvements in data quality. However, despite these advancements in multimodal generation, there remains a lack of a large-scale, high-quality image–text dataset specifically designed for medical illustration, which limits the application of such models in medical education and research.

Unified Understanding and Generation Models. Unified multimodal understanding and generation models have attracted significant attention from the research community. Existing approaches can be broadly categorized into three paradigms. (1) Autoregressive models [63, 8, 38, 48, 56, 59], which adopt a next-token prediction paradigm to generate both text and image tokens within a single unified model. (2) Models with additional diffusion modules [15, 64, 44, 58], which typically combine a pre-trained large language model (LLM) backbone with an external diffusion model. In these approaches, the LLM provides semantic conditions that guide the diffusion process for image generation. (3) Unified integrated transformer models [12, 33, 39], which tightly integrate LLMs and diffusion models within a single transformer architecture.

Reasoning in T2I Generation. Text-based reasoning models have demonstrated substantial progress in solving a wide range of real-world tasks [27, 20, 7]. More recently, interleaved reasoning has emerged as a promising paradigm that integrates non-text modalities into multi-step reasoning processes to address increasingly complex problems. In the context of text-to-image (T2I) generation, recent studies [16, 46, 26, 22, 43] investigate whether incorporating text-based reasoning or introspective mechanisms can improve the quality and reliability of generated images. These approaches primarily aim to enhance general-purpose image generation by embedding reasoning capabilities into unified multimodal models.

C Data Details

C.1 Construction of the Pretraining Data

We construct our pre-training corpus based on **BIOMEDICA** [37], a massive multimodal archive derived from the PubMed Central Open Access (PMC-OA) subset. The original dataset contains over **24 million** image-text pairs extracted from 6 million articles, featuring comprehensive metadata and expert-guided annotations.

To ensure visual consistency with medical illustrations, we implemented a *Coarse-to-Fine* filtering strategy:

1. **Taxonomy-Based Coarse Filtering:** Leveraging BIOMEDICA's 16 Global Taxonomies, we initially selected two categories most relevant to our target domain: *Illustrative Diagrams* and *Drawings*.
2. **VLM-Based Fine-Grained Filtering:** Despite the taxonomic selection, the initial subset retained significant noise irrelevant to medical illustration, such as administrative flowcharts, and low-quality sketches. To strictly purify the data, we employed a specialized visual selector. We manually annotated 6,000 samples to distinguish "Medical Illustrations" from generic imagery and used this data to fine-tune Qwen3-VL-8B-Instruct [2] as a binary classifier.

This pipeline distilled the massive corpus into a high-quality subset of **1.21 million** pairs, ensuring that our initialization phase focuses exclusively on semantically rich biomedical visualizations.

C.2 Construction of the SFT Data

C.2.1 Text-To-Image Data Filtering

To ensure the integrity and compliance of the dataset, we implemented a multi-stage filtering pipeline. This process sequentially addresses technical quality, copyright risks, and semantic relevance.

1. **Technical Quality Validation.** We first performed a validity check to discard corrupted image files that failed to decode. To strictly maintain the visual fidelity of the synthesized anatomy, we filtered out low-resolution samples, setting a threshold where the shortest edge of an image must exceed 512 pixels.
2. **Copyright and Source Filtering.** To strictly mitigate copyright risks, we employed a two-tiered filtering strategy. First, we screened the crawling metadata to identify images

[2] `https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct`

originating from known commercial stock photography websites and copyright-restricted domains. We aggressively excluded all data from these sources, prioritizing strict compliance even if it necessitated discarding high-quality anatomical illustrations. Second, we deployed a pre-trained watermark detection model to identify visual indicators of ownership. We adopted a zero-tolerance policy: rather than attempting to restore images via watermark removal or cropping, which risks artifact retention, any image detected with watermarks was completely discarded from the dataset.

3. **VLM-Based Semantic Filtering.** Data scraped from web engines often contains noise, such as generic clip art or irrelevant photography, despite keyword matching. To address this issue, we employ a vision–language model (VLM)–based fine-grained semantic filtering strategy. Specifically, we fine-tune **Qwen3-VL-8B-Instruct** as a binary classifier to distinguish professional medical illustrations from non-medical or low-quality visual content. The model is trained on a manually annotated subset, where images are labeled according to whether they satisfy strict medical illustration standards, including anatomical relevance, diagrammatic structure, and educational intent. This learned VLM filter enables robust removal of irrelevant images such as decorative graphics, generic diagrams, or illustrative noise, ensuring that only semantically valid medical illustrations are retained. By leveraging the same VLM configuration and prompt engineering, we rigorously verified that the visual content of each image strictly aligned with professional medical anatomical standards, discarding any data classified as non-medical or illustrative noise.

Through the above procedures, we obtain approximately **138K high-quality samples**.

C.2.2 Statistical Analysis of Text-To-Image Data

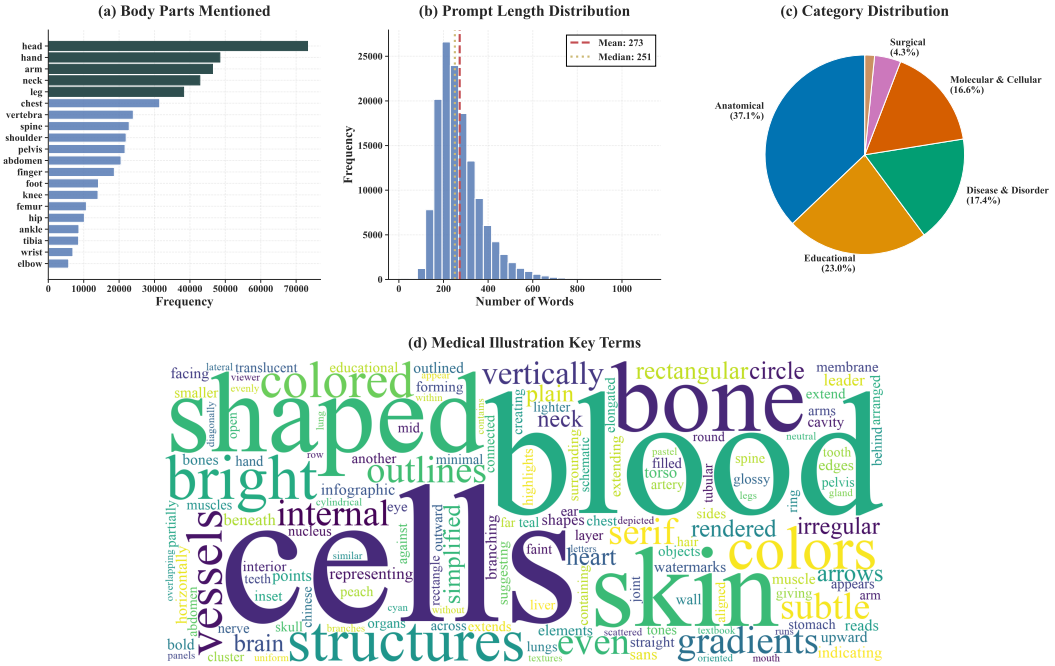


Figure 10: Statistical analysis of Stage II dataset, including (a) frequency distribution of anatomical terms, (b) prompt length distribution, (c) category composition, and (d) word cloud visualization of domain-specific terminology.

We present a comprehensive statistical analysis of the dataset metadata in Fig. 10. The frequency analysis of anatomical terms (Fig. 10a) reveals a strong emphasis on osteological structures, with terms such as *spine*, *vertebra*, and *femur* appearing most frequently, indicating that the dataset is heavily weighted toward musculoskeletal imagery. This trend is further corroborated by the word cloud visualization (Fig. 10d), which highlights dominant domain-specific vocabulary and confirms the prevalence of structural anatomical descriptors over pathological or procedural terms.

The distribution of prompt lengths (Fig. 10b) exhibits a pronounced right-skew, suggesting that while the majority of descriptions are relatively concise, a long tail of highly detailed prompts contributes substantially to the semantic richness of the corpus. In addition, the categorical breakdown (Fig. 10c) shows that *Anatomical* (37.1%) and *Educational* (23.0%) illustrations constitute the majority of the dataset, providing a strong foundation for learning canonical medical representations, whereas *Surgical* and *Other* categories are comparatively underrepresented.

C.3 Interleaved Data Construction

To enable introspection-aware refinement in medical illustration generation, we construct an interleaved text–image dataset through an explicit *generate–reflection–regeneration* pipeline. Unlike static text–image pairs, our data captures the full iterative reasoning trajectory, providing supervision signals for both visual generation and self-reflective correction.

Initial Image Generation. We first construct over 15,000 medical illustration seed prompts based on the Medical Subject Headings (MeSH) [3], a standardized and hierarchical biomedical ontology. These seed prompts systematically cover diverse medical concepts and serve as structured inputs for subsequent data generation. Given a medical illustration seed prompt T_{in} constructed based on MeSH concepts, we first generate a preliminary image $I^{(0)}$ using **BAGEL-Think**, a reasoning-enhanced text-to-image model. This initial output serves as a deliberately imperfect starting point, reflecting common failure modes such as anatomical omissions, structural inconsistencies, or partial instruction misalignment that are frequently observed in one-pass generation.

Multi-round Self-Reflection. Starting from $I^{(0)}$, we perform several rounds of self-reflection. At each round k, we employ strong vision–language models (VLMs), including **GPT-5.1**, to evaluate the image $I^{(k-1)}$ with respect to the original instruction T_{in}. The VLM outputs a structured textual assessment $T_{\text{ref}}^{(k)}$ that consists of two components: *(1) an explicit critique* of image quality, anatomical correctness, and instruction-following, and *(2) a concrete refinement instruction* describing how the image should be modified to better align with the medical intent of the prompt.

This reflection process emphasizes fine-grained, medically grounded reasoning, identifying specific visual elements that require correction rather than providing generic feedback.

Image Refinement. Conditioned on the refinement instruction $T_{\text{ref}}^{(k)}$, we then invoke image generation models capable of image editing, such as **Gemini-3-pro-image**, to modify the image from the previous round. This produces a refined image $I^{(k)}$ that explicitly incorporates the suggested corrections while preserving unrelated visual content. By iterating this process, we obtain a progressive refinement trajectory:

$$(T_{\text{in}}, I^{(0)}) \rightarrow T_{\text{ref}}^{(1)} \rightarrow I^{(1)} \rightarrow \cdots \rightarrow T_{\text{ref}}^{(K)} \rightarrow I^{(K)}.$$

Interleaved Data Collection. We collect both the textual reflections $\{T_{\text{ref}}^{(k)}\}$ and the corresponding visual states $\{I^{(k)}\}$ from each self-reflection loop. These interleaved text–image sequences form the supervision data for **Refinement SFT**, explicitly teaching the model how to: (i) evaluate its own generated images, (ii) articulate medically meaningful critiques, and (iii) apply corrective feedback to refine subsequent generations.

Quality Control. To ensure data reliability, we discard trajectories where reflections are vague, medically incorrect, or fail to produce measurable visual improvement after editing. Remaining samples are further reviewed to verify that each refinement step exhibits a clear causal relationship between the reflection and the resulting visual correction.

Overall, We construct over **12K interleaved reasoning data samples**.

[3] `https://www.nlm.nih.gov/mesh/meshhome.html`

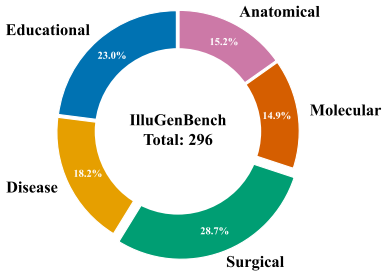
D Evaluation Details

D.1 IlluGenBenchmark Statistics

Table 7: Key statistics of IlluGenBench.

Statistics	Value
Minimum Number of Criteria	30
Maximum Number of Criteria	31
Average Number of Criteria	30.46
Minimum Criteria Length (words)	29.13
Maximum Criteria Length (words)	31.13
Average Criteria Length (words)	30.34
Minimum Prompt Length (words)	295
Maximum Prompt Length (words)	439
Average Prompt Length (words)	366

Figure 11: Category distribution of IlluGenBench.



IlluGenBench is constructed with carefully controlled prompt and rubric characteristics. As shown in Table 7, each prompt is associated with a fixed number of rubric criteria, ensuring consistent evaluation granularity across all samples. The criteria are also well-balanced in length, with an average of 30.34 words, indicating uniform semantic density and reducing bias introduced by overly concise or verbose descriptions. Prompt lengths range from 295 to 439 words, with an average of 366 words, reflecting the complexity required to specify fine-grained visual generation tasks. In addition, Figure 11 illustrates the category distribution of IlluGenBench, demonstrating broad coverage across diverse domains and supporting comprehensive evaluation of illustration generation models.

D.2 Experimental Setup

Training Details. We freeze the vision encoder to maintain a stable visual codebook, while fully fine-tuning the LLM backbone and all task-specific modules, including generation embeddings, prediction heads, and alignment projectors.

Table 8: Training settings for the three-stage.

Hyper-parameters	Stage I (PT)	Stage II (SFT)	Stage III (GRPO)
Optimizer	AdamW	AdamW	AdamW
LR	1.0e−4	1.0e−5	2.0e−5
LR scheduler	Cosine	Cosine	Linear + Cosine
Batch size	32	16	8
Training steps	293K	44.7K	3K
Warmup steps	800	800	300
Weight decay	0.05	0.05	0.05
Numerical precision	bfloat16	bfloat16	bfloat16
Resource usage	8 NVIDIA A800	8 NVIDIA A800	8 NVIDIA A800

Baseline Models. We select representative T2I generation models for comparison, including **(1) Closed-source commercial models** such as GPT-Image-1 [42] and Gemini-2.5-Flash-Image [19]; **(2) Open-source models** such as Qwen-Image [62] and HiDream-I1-Full [3]; **(3) Unified Models** such as Janus-Pro [9] and Janus-4o [6]; and **(4) Reasoning models** such as Uni-CoT [46]. For a fair comparison, we conducted inference only once, avoiding any cherry-picking of results. For all models, we use their default T2I configuration for inference.

MLLM-as-Judge Implementation. We employ Gemini-3-Pro (temperature is set to 0) as an automated judge. Two senior experts provide human evaluations for all 296 medical illustration tasks for consistency assessment.

Human Evaluation. To evaluate alignment with human preferences in medical illustration generation, we conducted a human study comparing MedIGen with Janus-Pro-7B, BAGEL, Show-o2, and

Table 9: List of evaluated T2I generation models and their links.

Model Name	Model Link
Commercial T2I Generation Models	
GPT-Image-1	https://platform.openai.com/docs/models/gpt-image-1
GPT-Image-1.5	https://platform.openai.com/docs/models/gpt-image-1.5
Gemini-2.5-Flash-Image	https://developers.googleblog.com/introducing-gemini-2-5-flash-image/
Gemini-3-Pro-Image-Preview	https://aistudio.google.com/models/gemini-3-pro-image
Seedream4.5	https://seed.bytedance.com/en/seedream4_5
Kling-Image-v2.1	https://app.klingai.com/global/
Open-Source T2I Generation Models	
Qwen-Image	https://huggingface.co/Qwen/Qwen-Image
Qwen-Image-2512	https://huggingface.co/Qwen/Qwen-Image-2512
FLUX.1-dev	https://huggingface.co/black-forest-labs/FLUX.1-dev
Stable-Diffusion-3.5-large	https://huggingface.co/stabilityai/stable-diffusion-3.5-large
Chroma1-HD	https://huggingface.co/lodestones/Chroma1-HD
HiDream-I1-Full	https://huggingface.co/HiDream-ai/HiDream-I1-Full
Lumina-Image-2.0	https://huggingface.co/Alpha-VLLM/Lumina-Image-2.0
Playground-v2.5	https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic
SDXL	https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
ERNIE-Image	https://huggingface.co/baidu/ERNIE-Image
Unified Understanding and Generation Models	
Janus-Pro-1B	https://huggingface.co/deepseek-ai/Janus-Pro-1B
Janus-Pro-7B	https://huggingface.co/deepseek-ai/Janus-Pro-7B
Janus-4o	https://huggingface.co/FreedomIntelligence/Janus-4o-7B
BAGEL	https://huggingface.co/ByteDance-Seed/BAGEL-7B-MoT
BLIP3o-NEXT	https://huggingface.co/BLIP3o/BLIP3o-Model-8B
UniWorld-V1	https://huggingface.co/LanguageBind/UniWorld-V1
Emu3.5	https://huggingface.co/BAAI/Emu3.5
Show-o2	https://huggingface.co/showlab/show-o2
GLM-Image	https://huggingface.co/zai-org/GLM-Image
T2I Reasoning Generation Models	
GoT	https://huggingface.co/LucasFang/GoT-6B
Janus-Pro-R1	https://huggingface.co/midbee/Janus-Pro-R1-7B
Uni-CoT (v0.2)	https://huggingface.co/Fr0zencr4nE/UniCoT-7B-MoT-v0.2
T2I-R1	https://huggingface.co/CaraJ/T2I-R1

FLUX.1-dev. For each task category, we randomly selected 20 tasks. Model outputs were randomly ordered, and evaluators provided an overall judgment by selecting one of three outcomes: win, loss, or tie. Results are reported as preference ratios.

D.3 Macro F1 of Human Consistency

Following HealthBench [1], we use *Macro F1* to evaluate the effectiveness of MLLM-based judges as proxies for human judgment under a binary classification setting. For each task and rubric criterion, we compare the judgments produced by the model judge with expert-annotated ground-truth labels, and compute F1 scores over the two classes {*Satisfied*, *Not Satisfied*}.

$$F_1 - score = \frac{2\,TP}{2\,TP + FP + FN}, \tag{2}$$

where TP, FP, and FN denote true positives, false positives, and false negatives. We analyze the key factors affecting the consistency between MLLM-based judges and human experts.

D.4 Access link for the Baseline model

D.5 Scoring Methodology

We evaluate model outputs using a rubric-based scoring method that operates at the *task level* and aggregates scores via *macro averaging*.

Task-Level Score. Each task (prompt) k is associated with a set of rubric criteria $C_k = \{r_1, \ldots, r_{|C_k|}\}$. Each criterion r_i is assigned a scalar weight $w_i \in \mathbb{R}$, where positive weights reward desirable behaviors and negative weights penalize failure modes.

For a given model response, a judge assigns a binary verdict to each criterion:

$$m_i = \begin{cases} 1, & \text{if criterion } r_i \text{ is satisfied,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The task-level score S_k is computed as a normalized weighted sum:

$$S_k = \frac{\sum\limits_{r_i \in C_k} w_i \, m_i}{\sum\limits_{r_i \in C_k,\, w_i > 0} w_i}. \tag{4}$$

The denominator normalizes by the sum of positive weights, corresponding to the maximum achievable score for task k. Negative-weight criteria contribute only to the numerator and therefore reduce the score when triggered. This normalization ensures that all task scores are comparable across tasks with different rubric sizes or weight distributions.

Average Score. Given a benchmark consisting of K tasks, we compute the overall model score as a simple macro average over task-level scores:

$$S_{\text{overall}} = \frac{1}{K} \sum_{k=1}^{K} S_k. \tag{5}$$

Each task contributes equally to the final score, preventing tasks with more rubric items or higher total weight from disproportionately influencing the evaluation.

Category-Level Scores. Rubric criteria are grouped into semantic categories (e.g., *Scientific Accuracy*, *Structural Correctness*, *Semantic Alignment*). Let $C_{k,c} \subseteq C_k$ denote the subset of criteria for task k belonging to category c.

For each task and category, we first compute a task-level category score:

$$S_{k,c} = \frac{\sum\limits_{r_i \in C_{k,c}} w_i \, m_i}{\sum\limits_{r_i \in C_{k,c},\, w_i > 0} w_i}, \quad \text{if} \sum_{r_i \in C_{k,c},\, w_i > 0} w_i > 0. \tag{6}$$

Let T_c denote the set of tasks that contain at least one criterion from category c. The category-level score is then computed as a macro average across tasks:

$$S_c = \frac{1}{|T_c|} \sum_{k \in T_c} S_{k,c}. \tag{7}$$

This formulation mirrors the overall score computation while restricting attention to a single rubric category, ensuring task-level fairness and avoiding rubric-level micro-averaging.

D.6 Information about the experts

Expert Panel. Five medical experts participated in the evaluation of the medical illustration generation task, contributing to both evaluation protocol design and model assessment. The panel consisted of **three junior experts (E1–E3)**, each holding an MD or equivalent clinical degree with 3–4 years of post-graduate training, and **two senior experts (E4–E5)**, board-certified physicians with over 6 years of clinical and academic experience. All experts have prior exposure to medical imaging or anatomical illustration review. To preserve privacy, all identities are anonymized.

Table 12: Performance comparison of different T2I generation models on IlluGenBench across five task categories. Bold indicates the best performance in each column.

Model	Educational↑	Anatomical↑	Surgical↑	Disease↑	Molecular↑
Commercial T2I Generation Models					
GPT-Image-1	0.8015	0.8561	0.8433	0.8300	0.8484
GPT-Image-1.5	0.8297	0.8420	0.8478	0.8167	0.8494
GPT-Image-2	0.7504	0.8000	0.8056	0.7659	0.7502
Gemini-2.5-Flash	0.7063	0.7580	0.7001	0.7227	0.8225
Gemini-3-Pro	**0.8644**	**0.8803**	**0.8619**	**0.8683**	**0.9045**
Gemini-3.1-Flash	0.8344	0.8794	**0.8761**	**0.8747**	0.8919
Doubao-Seedream-4.5	0.7437	0.8320	0.7504	0.7396	0.8080
Kling-Image-v2.1	0.1942	0.1931	0.1911	0.2490	0.0947
Open-Source T2I Generation Models					
ERNIE-Image	0.7273	0.8064	0.7713	0.7568	0.8384
Qwen-Image	0.4050	0.5793	0.4184	0.4444	0.3316
Qwen-Image-2512	0.5482	0.6854	0.5676	0.5924	0.6619
FLUX.1-dev	0.3718	0.4764	0.3806	0.3964	0.3486
Stable-Diffusion-XL	0.0910	0.1839	0.1139	0.1463	0.0197
Stable-Diffusion-3.5	0.1968	0.3150	0.2022	0.2720	0.0825
Chroma1-HD	0.3946	0.4343	0.4406	0.4517	0.3548
HiDream-I1	0.2181	0.3941	0.2309	0.2975	0.1749
Lumina-Image-2.0	0.2932	0.3542	0.2972	0.3653	0.2930
Playground-v2.5	0.0735	0.1290	0.0859	0.1063	0.0239
Unified Understanding and Generation Models					
Janus	0.3257	0.3424	0.3254	0.3344	0.3161
Janus-1B	0.1914	0.2133	0.2030	0.2730	0.2231
Janus-4o-7B	0.3928	0.4224	0.4230	0.4636	0.4904
BAGEL	0.3866	0.3591	0.3870	0.4006	0.4233
BLIP3o	0.3208	0.4135	0.3111	0.3903	0.3111
GLM-Image	0.4719	0.5590	0.4194	0.4661	0.6203
T2I Reasoning Models					
GoT	0.2538	0.2694	0.2416	0.2649	0.2593
Janus-Pro-R1	0.0878	0.0583	0.0380	0.0759	-0.0112
Uni-CoT (v0.2)	0.4267	0.4342	0.3975	0.4113	0.3958
T2I-R1	0.3032	0.3051	0.2657	0.3003	0.2822
Our Models					
MedIGen	0.7021	0.5558	0.5777	0.4605	0.8253

Table 10: Junior Experts

Expert ID	Years of Practice
E1	3
E2	3
E3	4

Table 11: Senior Experts

Expert ID	Years of Practice
E4	6
E5	7
-	-

E More Analysis Results

E.1 Evaluation Results by Task Category

Table 12 reports the performance of representative text-to-image (T2I) generation models on IlluGenBench, evaluated across five different categories: Educational, Anatomical, Surgical, Disease,

and Molecular. Commercial models consistently outperform open-source and unified understanding–generation models, with Gemini-3-Pro achieving the best results in all five task categories, indicating strong overall medical fidelity and domain coverage. Among open-source methods, Qwen-Image-2512 and Chroma1-HD demonstrate comparatively stronger anatomical and disease-level understanding, though a substantial performance gap remains relative to commercial systems. Unified models exhibit moderate and more balanced performance, suggesting improved cross-modal reasoning but limited generation precision. Notably, MedIGen attains competitive scores in the Educational and Molecular categories, narrowing the gap with leading commercial models and highlighting its effectiveness in specialized medical knowledge representation.

E.2 Analysis of Maximum Generation Rounds

MedIGen provides a simple inference-time hyperparameter, maximum generation rounds (MGR), to control the trade-off between computation and generation quality. MGR$=1$ means that the model performs the initial generation and self-reflection but does not invoke additional re-generation. MGR$=3$ is our default setting, allowing the model to iteratively revise difficult cases when self-reflection detects semantic or anatomical errors.

Table 13: **Effect of maximum generation rounds on IlluGenBench.** MGR controls the maximum number of image generation attempts during inference.

Model	Sci. Acc.	Struct. Corr.	Sem. Align.	Avg.
Janus-Pro-7B	0.2977	0.2244	0.4628	0.3276
MedIGen (MGR$=1$)	0.5743	0.5082	0.6831	0.5853
MedIGen (MGR$=3$)	0.6064	0.5368	0.7111	0.6183

The results show that MedIGen remains substantially stronger than Janus-Pro-7B even under MGR$=1$, improving the average score from 0.3276 to 0.5853 without additional re-generation rounds. Increasing MGR to 3 further improves the average score to 0.6183, indicating that iterative refinement mainly benefits harder prompts while allowing users to adjust inference cost according to deployment requirements.

E.3 Expert Audit of Disease Failure Cases

To identify where MedIGen fails on pathologically complex prompts, we conduct a manual expert audit of 56 Disease-category cases. We separate each trajectory into two stages: whether self-reflection correctly identifies the error, and whether re-generation successfully implements the correction.

Table 14: **Expert audit of Disease-category failures.** The audit separates failures in self-reflection from failures in re-generation.

Phase	Process Outcome	Case Count
Self-Reflection	Successful Identification	34/56
Self-Reflection	Missed Detection or No Error	22/56
Re-Generation	Failed to Correct (Round 1)	29/34
Re-Generation	Failed After 2+ Rounds	19/29

The audit shows that MedIGen can detect a substantial portion of disease-related errors: self-reflection successfully identifies issues in 34 out of 56 cases. However, among these identified cases, 29 out of 34 still fail to produce a correct image after the first re-generation round, and 19 of these 29 remain incorrect after multiple rounds. This indicates that the dominant bottleneck is not only recognizing pathological inaccuracies, but converting the textual critique into an anatomically and pathologically correct visual correction. Disease illustrations often require modeling subtle deviations from normal anatomy, lesion morphology, and spatial relationships between pathological and healthy tissues.

E.4 Cross-Evaluator Consistency Analysis

A key concern in RL from AI feedback is *reward hacking*: the model may learn to exploit biases in the reward model rather than genuinely improving on the target objective. We use Qwen3-VL-30B-A3B as the reward model during Stage 3 training. To verify that MedIGen learns generalizable medical accuracy rather than overfitting to Qwen's specific preferences, we conduct a cross-evaluator consistency analysis.

We compare MedIGen-SFT (trained through Stages 1–2 only) and MedIGen-RL (trained through all three stages) on the complete IlluGenBench test set (296 tasks). We evaluate both models using three independent judges: (1) Qwen3-VL-30B-A3B (the training reward model), (2) Gemini-3-Pro (an independent MLLM judge not used during training), and (3) human medical experts (three board-certified physicians). All MLLM judges use the same IlluGenBench rubric evaluating Scientific Accuracy, Structural Correctness, and Semantic Alignment.

Table 15 shows that all three independent evaluators observe consistent improvements from the RL stage. If MedIGen were merely learning to exploit Qwen's biases, we would expect improvements only on Qwen's evaluations, with flat or degraded performance on Gemini and human judgments. Instead, the gains generalize across all evaluators, including those never seen during training. The largest improvement appears on Qwen (+0.07), which is expected since the model directly optimizes for this signal. However, Gemini (+0.03) and human experts (+45pp preference) also show substantial gains, indicating that the RL stage captures genuine improvements in medical accuracy.

Table 15: Cross-evaluator consistency analysis. All three independent evaluators observe consistent improvements from the RL stage, demonstrating that the model learns generalizable medical accuracy rather than overfitting to the reward model's biases.

Evaluator	**Role**	**SFT**	**RL**	Δ	p**-value**
Qwen3-VL-30B-A3B	Training reward	0.5743	0.6436	+0.0693	< 0.001
Gemini-3-Pro	Independent judge	0.5885	0.6183	+0.0298	< 0.001
Human experts	Gold standard	19% prefer	64% prefer	+45pp	< 0.001

Human evaluation: 100 tasks, 3 experts per task, majority vote. Tie rate: 17%.

Table 16 breaks down improvements by evaluation dimension. The RL stage consistently improves all three dimensions across both MLLM judges, with the largest gains in Scientific Accuracy. Table 17 shows high inter-evaluator correlations (Pearson $r > 0.78$), demonstrating that different judges largely agree on what constitutes high-quality medical illustration. This shared understanding is crucial: RL training relies on *relative* comparisons rather than absolute scores. As long as the reward model's ranking aligns with human preferences, it provides a valid learning signal.

Table 16: Per-dimension performance breakdown. The RL stage improves all three evaluation dimensions consistently across both MLLM judges.

Evaluator	**Dimension**	**SFT**	**RL**	Δ
Qwen3-VL-30B-A3B	Scientific Accuracy	0.5743	0.6521	+0.0778
	Structural Correctness	0.5082	0.5736	+0.0654
	Semantic Alignment	0.6831	0.7351	+0.0520
Gemini-3-Pro	Scientific Accuracy	0.6064	0.6389	+0.0325
	Structural Correctness	0.5368	0.5642	+0.0274
	Semantic Alignment	0.7111	0.7318	+0.0207

Table 17: Inter-evaluator agreement. High correlations indicate shared understanding of medical illustration quality across different judges.

Evaluator Pair	**Pearson** r	**Spearman** ρ
Qwen $\leftrightarrow$ Gemini	0.823	0.817
Qwen $\leftrightarrow$ Human	0.781	0.769
Gemini $\leftrightarrow$ Human	0.794	0.786

Beyond judge scores, we observe that the RL stage increases the first-pass success rate from 64% to 80%. This metric—the fraction of prompts where the model generates a satisfactory image without requiring self-reflection and regeneration—is not directly observable to the reward model during training. Its improvement suggests that the model develops a more robust internal understanding of medical illustration requirements, rather than simply learning to manipulate the reward signal.

F Prompt Details

F.1 Prompt used for Caption Generation

Prompt used for Caption generation.

{Medical Illustration}
Metadata: {Metadata}
Your current task is: Based on the medical illustration and metadata I provide, generate a detailed, objective, and precise caption focused on anatomical and biological structures. This caption will be used as a prompt for a text-to-image model in order to replicate the original medical illustration as closely as possible.
Please follow these requirements:

1. Anatomical Description: Provide a detailed and accurate description of the organs, tissues, cells, or body parts shown. Mention their structure, orientation, and any visible biological processes.
2. Visual elements: Include colors, lighting, texture, materials, and specific medical illustration styles (e.g., cross-sections, 3D medical rendering, diagrammatic, histological view).
3. Spatial relationships: Explain the layout and relative positions of various anatomical components within the frame.
4. Ignore Textual Content: STRICTLY IGNORE IGNORE any text labels, pointers with letters/numbers, watermarks, or legends present in the image. Do not describe or mention them. Focus only on the physical structures.
5. Avoid subjective words: Do not use evaluative terms such as "beautiful" or "clear." Only use objective anatomical and visual descriptions.
6. Completeness: Ensure the caption is sufficiently detailed so that a text-to-image model can accurately reconstruct the biological structures even without seeing the original image.

Final output format: A single, complete English caption (suitable for direct input into a text-to-image model), without any extra explanation.

F.2 Prompt used for Rubric Generation

Prompt used for Rubric Generation.

[System Prompt]
You are an expert rubric writer specializing in medical and anatomical illustration. Your task is to generate a highly granular, concise, and professional evaluation rubric for medical illustrations.

[User Prompt]
You are an expert rubric writer. Your task is to generate a self-contained set of evaluation criteria ("rubrics") for judging how good a medical illustration is based on a given Final Prompt and the illustration itself.
Rubrics MUST cover the following three dimensions:

- Scientific Accuracy: Evaluate individual anatomical or biological facts (e.g., Is the heart's left ventricle correctly labeled?).
- Structural Correctness: Evaluate individual spatial relationships, proportions, and connectivity (e.g., Does the pulmonary artery correctly branch behind the aorta?).
- Semantic Alignment: Verify if specific keywords and scene descriptions from the Final Prompt are precisely visualized (e.g., Is the blue lighting mentioned in the prompt accurately reflected?).

Each rubric item must be self-contained; non-expert readers should not need to infer missing information or consult external sources.

Required Category Prefixes
Each rubric description must begin with one of the following category prefixes:

- Essential Criteria:
- Important Criteria:
- Optional Criteria:
- Pitfall Criteria:

Inputs
The rubric generation process is based on the following inputs:

- Final Prompt: {final_prompt}
- Image: The provided medical illustration.

Total Number of Rubric Items
For each of the three dimensions (scientific_accuracy, structural_correctness, and semantic_alignment), generate between 5 and 10 rubric items. This results in a total of 15 to 30 items per illustration.

Structure of Each Rubric Item
Each rubric item must contain the following elements:

- description: One sentence that begins with the appropriate category prefix and explicitly states exactly what the evaluator should look for in the image.
- weight:
  - Essential, Important, Optional: integer from 1 to 5, where 5 is the most important.
  - Pitfall: negative integer, either -1 or -2.

Category Guidance

1. Essential Criteria:
- Critical anatomical facts, safety-related visual elements, or core requirements of the prompt.
- If missing or incorrect, the illustration is scientifically invalid.
- Typical weight: 5.

2. Important Criteria:
- Key structural details, labeling clarity, or overall completeness.
- Strongly affects the educational or clinical quality of the illustration.
- Typical weight: 3 to 4.

3. Optional Criteria:
- Stylistic enhancements, aesthetic details, or additional depth.
- Nice to have, but not deal-breaking.
- Typical weight: 1 to 2.

4. Pitfall Criteria:
- Common anatomical mistakes, misleading visual cues, or omissions specific to the prompt.
- Each description must begin with one of the following forms:
  - Pitfall Criteria: Does not show ...
  - Pitfall Criteria: Incorrectly represents ...
- Weight must be -1 or -2.

Self-Containment Requirements
To ensure clarity and independence:

- Explicitly reference specific anatomical structures or features mentioned in the prompt.
- Avoid vague phrasing such as “the correct part” without specification.
- If a specific label is required, state it explicitly (e.g., Includes a clear label for the Left Ventricle).
- If a spatial relationship is critical, state it explicitly (e.g., Shows the pulmonary artery branching correctly behind the aorta).

Output Format
The final output must be a JSON object with three keys corresponding to the dimensions scientific_accuracy, structural_correctness, and semantic_alignment.

- Each key contains a list of rubric objects.
- Each rubric object must contain exactly two keys: description and weight.
- Do not include extra keys such as title.
- Do not copy large blocks of the prompt verbatim.

Example structure:

```
{
  "scientific_accuracy": [
    { "description": "Essential Criteria: ...", "weight": 5 }
  ],
  "structural_correctness": [ ... ],
  "semantic_alignment": [ ... ]
}
```

F.3 Prompt used for Rubric-Based Image Evaluation

Prompt used for Rubric-Based Image Evaluation.

[System Prompt]
You are an expert evaluator tasked with assessing whether a generated image satisfies specific rubric criteria. Your evaluation must be precise, objective, and based solely on the evidence present in the image.

Evaluation Framework
You will evaluate each rubric criterion using a binary satisfaction scale:

- **Not Satisfied (Score: 0.0)**: The image fails to meet the criterion. Key elements are missing, incorrect, or inadequately addressed.
- **Satisfied (Score: 1.0)**: The image meets the criterion. All required elements are present and correctly depicted.

Evaluation Process

1. **Understand the Criterion**: Carefully read and interpret what each rubric item is asking for.
2. **Search for Evidence**: Systematically review the generated image for relevant visual content that addresses the criterion.
3. **Assess Completeness**: Evaluate whether the evidence fully satisfies the criterion.
4. **Provide Reasoning**: Explain your evaluation with specific references to visual elements in the image.

Important Guidelines

- Base your evaluation **ONLY** on what is explicitly present in the generated image.
- Do not make assumptions about implied or missing content.
- Be consistent in your evaluation standards across all criteria.
- Provide specific examples from the image to support your verdict.

[User Prompt]
Visual Content
You are provided with two images. Please pay close attention to their order:

1. The **FIRST** image is the **Original Reference Image**. This serves as the ground truth.
2. The **SECOND** image is the **Generated Image to Evaluate**.

Rubric Criteria to Evaluate
Category: {category}
{criteria_text}

Your Task
Evaluate whether the **SECOND** image (Generated Image) satisfies **EACH** of the above rubric criteria. You should use the **FIRST** image (Original Reference) to understand the expected visual elements and ground truth.

Required Response Format
Provide your evaluation in the following JSON format:

```
{
  "evaluations": {
    "criterion_id_here": {
      "verdict": "[Not Satisfied/Satisfied]",
      "score": [0.0/1.0],
      "confidence": [0.0-1.0],
      "reasoning": "Detailed explanation with specific evidence from the image",
      "evidence_description": ["Description of visual element 1", ...],
      "missing_elements": ["Element 1 that would improve satisfaction", ...]
    }
  }
}
```

Important: The keys in `"evaluations"` must exactly match the string representation of the Criterion IDs provided above (e.g., "1", "2", "3", etc.). Ensure the response contains **ONLY** the JSON object, with no additional text.

G Examples of IlluGenBench

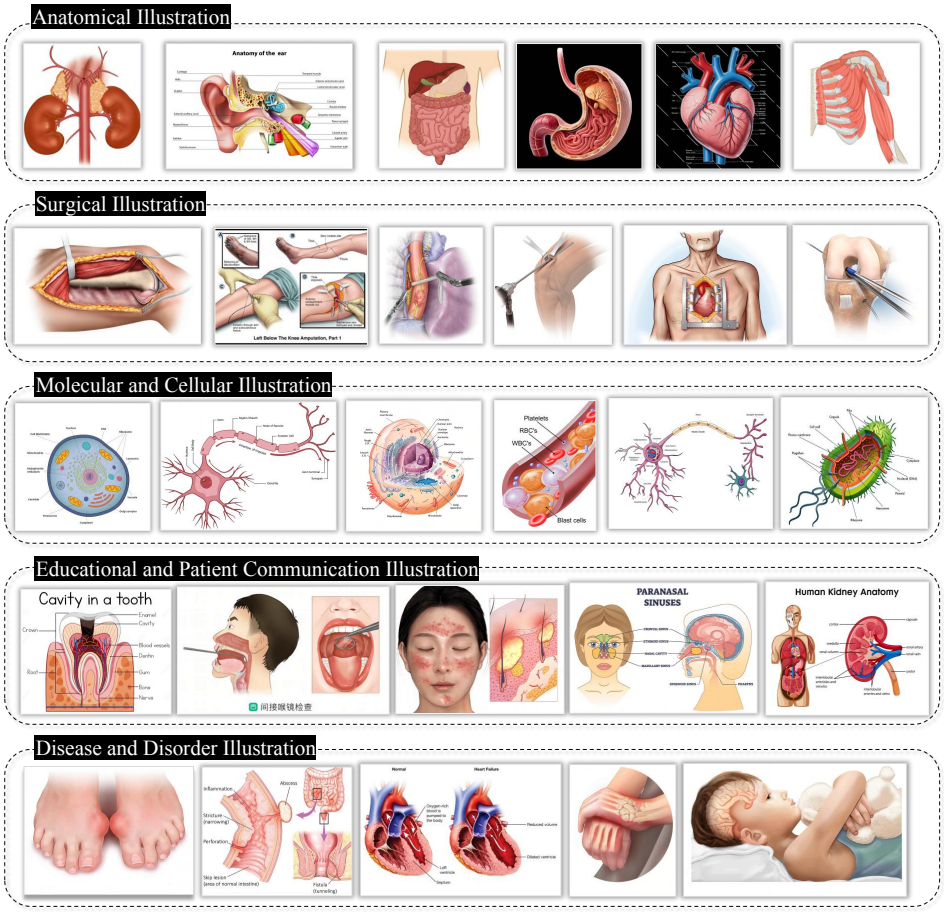


Figure 12: Examples of illustration data from different categories in IlluBench.

H Qualitative Comparison of MedIGen with Other Models

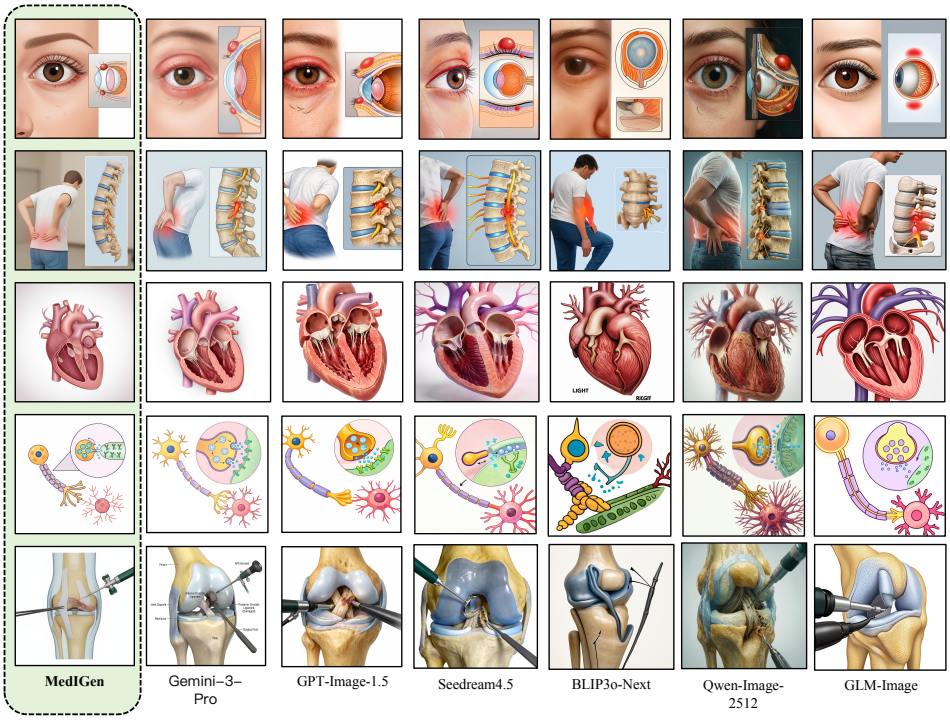


Figure 13: Qualitative Comparison of MedIGen with Other Models.

I Failure cases of MedIGen

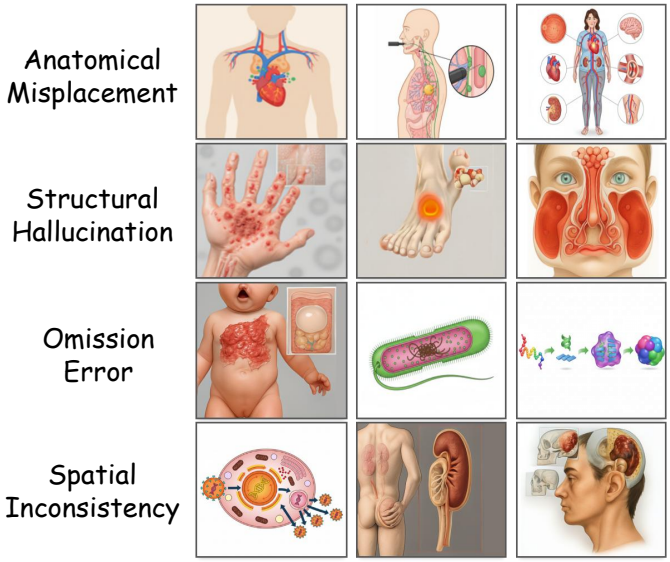


Figure 14: Failure cases of MedIGen in medical illustration generation.

As shown in Figure 14, the dominant failure modes include anatomical misplacement, structural hallucination, and spatial inconsistency. These errors highlight the need for stronger integration of precise biomedical relational knowledge to improve structural fidelity.

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: See abstract and introduction.

   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: See Appendix A.

   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: This paper does not present formal theoretical results requiring assumptions or proofs. Our contributions are empirical, focusing on model design, dataset construction, and experimental evaluation.

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: See section 5.1.

   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: All resources will be open-sourced to facilitate future research.

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: See section 5.1.

   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: We report single-run results for the main benchmark and do not systematically provide error bars or confidence intervals for all experiments, although selected analyses include significance statistics in the Appendix.

   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: See section 5.1.

   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: We reviewed the NeurIPS Code of Ethics and, to the best of our knowledge, the research conforms to it.

   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: See Appendix A.

    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: We implement responsible-release safeguards including copyright/source filtering, watermark-based exclusion, semantic quality filtering, benchmark–training set decontamination, and an explicit statement that the system is an assistive research tool rather than a clinical reference.

    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: We credit all third-party assets through citations and links, and we carefully follow their publicly stated licenses and terms of use, including additional filtering of restricted or copyright-sensitive sources during data collection.

    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: The new benchmark, data construction pipeline, and model details are documented in the paper and appendix, and the released assets will be accompanied by usage instructions and relevant metadata.

    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: [N/A]

    Justification: This paper does not involve crowdsourcing; the human evaluation is conducted by expert annotators as part of the research team, so no crowdworker instructions, screenshots, or compensation details apply.

    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: This paper does not involve crowdsourcing or human-subject research requiring IRB review, as the human evaluation was conducted by expert annotators as part of the research team.

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: We explicitly describe the use of LLMs/MLLMs in the core methodology, including the unified model backbone, interleaved data construction, rubric generation, automated judging, and reward-based training.

    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.