Med-Banana-50K: A Cross-modality Large-Scale Dataset for Text-guided Medical Image Editing

Zhihui Chen, Mengling Feng*
National University of Singapore
zhihui.chen@u.nus.edu

Abstract

Recent advances in multimodal large language models have enabled remarkable medical image editing capabilities. However, the research community's progress remains constrained by the absence of large-scale, high-quality, and openly accessible datasets built specifically for medical image editing with strict anatomical and clinical constraints. We introduce **Med-Banana-50K**, a comprehensive 50K-image dataset for instruction-based medical image editing spanning three modalities (chest X-ray, brain MRI, fundus photography) and 23 disease types. Our dataset is constructed by leveraging Gemini-2.5-Flash-Image to generate bidirectional edits (lesion addition and removal) from real medical images. What distinguishes Med-Banana-50K from general-domain editing datasets is our systematic approach to medical quality control: we employ LLM-as-Judge with a medically grounded rubric (instruction compliance, structural plausibility, realism, and fidelity preservation) and history-aware iterative refinement up to five rounds. Beyond single-turn editing, Med-Banana-50K includes 37K failed attempts with full conversation logs for preference learning and alignment research. By providing this large-scale, medically validated, and fully documented resource, Med-Banana-50K establishes a foundation for training and evaluating the next generation of medical image editing models. Our dataset and code are publicly available at [https://github.com/richardChenzhihui/med-banana-50k].

1 Introduction

Recent advances in multimodal large language models (MLLMs) such as GPT-4o Hurst et al. [2024] and Gemini-2.5-Flash-Image Comanici et al. [2025], along with diffusion-based editing models Brooks et al. [2023], Zhang et al. [2024], have demonstrated remarkable capabilities in text-guided image editing. These models can transform images based on natural language commands, from simple color adjustments to complex compositional changes.

Despite these advances, open research in medical image editing remains limited by the lack of large-scale, high-quality, and fully shareable datasets. Existing general-domain editing datasets Qian et al. [2025], Hui et al. [2024] often exhibit domain shifts when applied to medical imaging, and existing medical datasets focus primarily on diagnosis rather than editing. Medical image editing presents unique challenges: edits must preserve modality-specific noise and texture, maintain anatomical plausibility, enforce counterfactual minimal change (modifying only disease-relevant regions), and navigate complex licensing and privacy constraints.

To address these challenges, we introduce **Med-Banana-50K**, a comprehensive dataset of approximately 50K high-quality medical image edits built from real clinical images spanning three modalities (chest X-ray, brain MRI, fundus photography) and 23 disease types. Our dataset represents a systematic effort to create large-scale training data for instruction-based medical image editing that is both diverse and fully shareable under clear licensing terms.

Preprint. Under review.

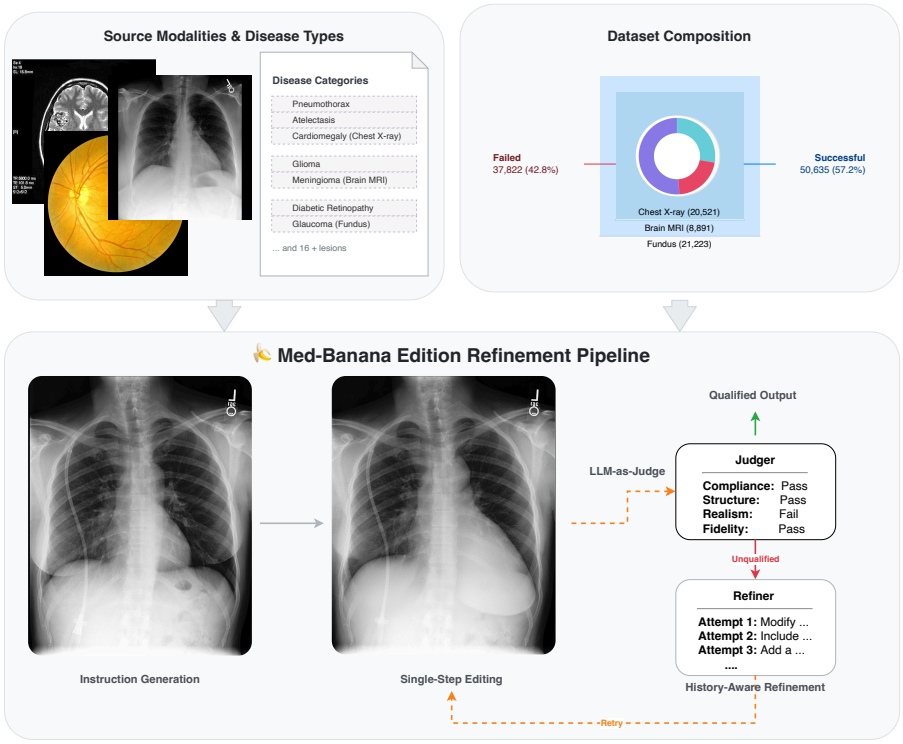

Figure 1: Pipeline: instruction generation, single-step editing, LLM-as-Judge evaluation, and history-aware refinement under fidelity, negative rules, and minimal change

Figure 1 illustrates our systematic approach to dataset construction. We leverage Gemini-2.5-Flash-Image to generate bidirectional edits (lesion addition and removal), employ Gemini-2.5-Pro as an automated judge for quality assurance through multi-dimensional medical scoring (instruction compliance, structural plausibility, realism, and fidelity preservation), and implement history-aware iterative refinement up to five rounds. Failed editing attempts are automatically retried and preserved as negative examples for preference learning, while successful edits form our core training data.

Our contributions are summarized as follows:

1. Large-scale shareable dataset: We release Med-Banana-50K, containing 50,635 high-quality medical image editing examples built from real clinical images, systematically organized by disease type and modality, with rigorous quality control through automated LLM-as-Judge scoring and multi-round refinement.

2. Multi-objective training support: Beyond the 50K successful single-turn examples, we provide 37K failed attempts with full conversation logs for preference learning and alignment methods like DPO Rafailov et al. [2024] and reward modeling, enabling research on robustness and iterative refinement.

3. Medical domain adaptation: We introduce a medically grounded quality rubric emphasizing fidelity preservation (noise/grain/artifacts), structural plausibility (anatomical correctness), and counterfactual minimal change, addressing unique constraints not covered by general-domain editing datasets.

2 Dataset Construction

We construct Med-Banana-50K through a systematic pipeline designed to ensure both scale and quality while respecting medical domain constraints. Our approach leverages state-of-the-art multimodal models for generation and evaluation while maintaining strict quality control at each stage.

We begin by describing our source images and medical editing taxonomy (Section 2.1), then detail our instruction generation procedure (Section 2.2), and present the construction of our single-turn dataset with automated quality assessment (Section 2.3).

2.1 Overview and Medical Editing Taxonomy

Our dataset is built upon images from three established medical imaging sources: MIMIC-CXR Johnson et al. [2019] for chest X-rays, Brain Tumor MRI Dataset Nickparvar [2021] for brain scans, and ODIR-5K Shangguan et al. [2019] for fundus photography. We organize medical image edits into a comprehensive taxonomy covering 23 disease types across 3 modalities and 2 editing directions (lesion addition and removal). Table 1 presents our complete taxonomy organized by modality and disease type. Our dataset is built upon images from three established medical imaging sources: MIMIC-CXR Johnson et al. [2019] for chest X-rays, Brain Tumor MRI Dataset Nickparvar [2021] for brain scans, and ODIR-5K Shangguan et al. [2019] for fundus photography. We organize medical image edits into a comprehensive taxonomy covering 23 disease types across 3 modalities and 2 editing directions (lesion addition and removal). Table 1 presents our complete taxonomy organized by modality and disease type.

Modality coverage. Our dataset spans three major medical imaging modalities: (1) *Chest X-ray* (12 pathology types): frontal radiographs covering common thoracic conditions including pneumothorax, pleural effusion, atelectasis, consolidation, edema, cardiomegaly, fracture, and lung lesions; (2) *Brain MRI* (4 tumor types): T1-weighted or FLAIR sequences showing glioma, meningioma, pituitary tumor, and no-tumor cases; (3) *Fundus photography* (7 disease types): color retinal images depicting diabetic retinopathy, glaucoma, age-related macular degeneration, cataract, hypertension, myopia, and normal cases.

Bidirectional editing tasks. For each (image, disease) pair, we construct two complementary editing directions. *Lesion addition* (normal $\rightarrow$ diseased) synthesizes pathological findings in healthy images, enabling data augmentation and counterfactual generation. *Lesion removal* (diseased $\rightarrow$ normal) simulates therapeutic outcomes or generates healthy references for comparison. This bidirectional design yields 6 distinct task combinations (3 modalities $\times$ 2 directions).

Image preprocessing. All images are resized to 1024$\times$1024 resolution and converted to JPEG format (quality 95) to balance file size and perceptual quality. We preserve original metadata where available for traceability but remove all patient identifiers from released data.

2.2 Instruction Generation

For each (image, disease, task) tuple, we generate an English editing instruction using Gemini-2.5-Pro with a carefully designed system prompt. The prompt emphasizes three requirements: (1) *Understandability*: instructions should be comprehensible to non-medical personnel while remaining medically plausible; (2) *Specificity*: instructions must reference visible anatomical landmarks and disease characteristics; (3) *Naturalness*: phrasing should resemble how a user might naturally request an edit.

Critically, the instruction generation prompt enforces our three core medical constraints: **fidelity preservation** (preserve noise, grain, device artifacts, imaging characteristics), **negative rules** (no text/labels, no sharp unnatural edges, no repetitive structures, no non-target pathologies), and **counterfactual minimality** (modify only disease-relevant regions). An example instruction for pneumothorax addition might be: "Add signs of pneumothorax in the right lung field, showing a visible visceral pleural line with absence of lung markings beyond it, while preserving the original image grain and contrast."

2.3 Single-Turn Medical Image Editing

Each edit instruction is executed by Gemini-2.5-Flash-Image, a state-of-the-art multimodal editing model. After generating an edit, Gemini-2.5-Pro serves as an automatic judge that evaluates the edit quality and determines whether it should be retained in the dataset.

extbfJudging process. The judge evaluates edits using four medical-specific criteria with weighted scoring: *Instruction Compliance* (40%), which measures how well the edit fulfills the prompt (e.g.,

Table 1: Dataset composition summary by modality and task. Success rate = successful / (successful + failed). Average rounds computed only for successful tasks.

Modality	Task	Diseases	Success	Failed	Avg. rounds
Chest X-ray	Add	12	9,854	7,971	1.42
Chest X-ray	Remove	12	10,667	4,750	1.28
Brain MRI	Add	4	4,536	8,630	1.35
Brain MRI	Remove	4	4,355	6,949	1.19
Fundus	Add	7	18,505	3,162	1.51
Fundus	Remove	7	2,718	6,360	1.33
Total	–	**23+**	**50,635**	**37,822**	**1.35**

is the target pathology clearly visible for addition tasks?); *Structural Plausibility* (25%), which checks for anatomically correct and medically plausible disease presentations; *Realism* (20%), which assesses natural integration without visible artifacts or blending issues; and *Fidelity Preservation* (15%), which ensures modality-specific noise/grain, device artifacts, and minimal intervention are maintained. We provide the complete judge prompt in the supplemental materials. The resulting score is aggregated into a single quality metric; images with scores above a strict threshold (empirically set to $\sim$0.7) are labeled as successful edits, while those below are categorized as failures.

Iterative refinement with history awareness. If the judge deems an edit unqualified, we initiate a refinement loop (maximum 5 rounds). Crucially, we employ **history-aware prompt updating**: the refinement step receives (1) the original image, (2) complete history of all previous instruction variants and their corresponding judge verdicts, and (3) a meta-instruction to analyze failure patterns and generate an improved instruction that avoids repeated mistakes. Each refinement round generates a new instruction and re-executes editing and judging from the original image—failed attempts do not accumulate, ensuring each iteration tests an independent instruction variant. This design prevents cyclic failures common in single-round feedback systems. If any round succeeds, the task terminates and the successful edit is saved. If all 5 rounds fail, the task is marked as failed, but all intermediate attempts and conversations are preserved as negative examples for preference learning.

Dataset outcomes. This self-evaluation process enables Med-Banana-50K to scale automatically while maintaining high semantic fidelity and visual realism:

- **Successful edits** ($\sim$50K) constitute the main dataset for supervised fine-tuning;
- **Failure cases** ($\sim$37K) are retained as negative examples paired with successful edits for preference learning and alignment research (e.g., DPO, reward modeling).

3 Dataset Composition and Statistics

Med-Banana-50K comprises three source datasets processed through bidirectional editing. Table 1 summarizes the scale, success rates, and round distributions across modalities and tasks. Overall, we generated 50,635 successful edits from 88,457 editing attempts (57.2% overall success rate), preserving 37,822 failed attempts with full conversations for analysis.

Modality-specific observations. Chest X-ray exhibits strong performance with 61.7% overall success rate (20,521 successful from 33,242 attempts), benefiting from MIMIC-CXR's large scale and clear pathology patterns. Brain MRI shows the lowest success rate (36.3%, 8,891/24,470), likely due to subtle tumor boundaries and complex anatomical structures that frequently trigger fidelity preservation failures. Fundus imaging achieves the highest success rate (69.0%, 21,223/30,745), as color fundus pathologies (diabetic retinopathy, glaucoma) have well-defined visual patterns that editing models can reliably reproduce.

Task asymmetry. Removal tasks generally require fewer refinement rounds than addition tasks (1.27 vs. 1.43 rounds on average), suggesting that erasing pathology while preserving structure is more straightforward than synthesizing realistic disease patterns. However, removal tasks also exhibit higher failure rates for subtle pathologies, where the judge cannot confidently verify complete removal.

extbfDisease difficulty variation. Within chest X-ray, Fracture addition shows the lowest success rate (15.8%), often failing due to the judge detecting unnatural edges or insufficient fracture lines. Consolidation addition also struggles (47.9% success), as subtle infiltrates are difficult to balance between visibility and naturalness. By contrast, Pneumothorax and Enlarged Cardiomediastinum have highly recognizable patterns and succeed more frequently (68.5% and 78.9%). Figure 3 visualizes the per-disease success and failure counts and shows the rounds-to-success distribution: 58% of successful tasks pass in round 1, 25% in round 2, and the remaining 17% require 3–5 rounds.

4 Quality Control with LLM-as-Judge

A key innovation of Med-Banana-50K is scalable, medically grounded quality control through LLM-as-Judge. Traditional human evaluation is costly and does not scale to tens of thousands of edits; rule-based metrics (SSIM, LPIPS, FID) fail to capture anatomical plausibility or instruction compliance. We leverage Gemini-2.5-Pro configured with deliberate reasoning (*thinking mode*) to perform structured evaluation along four weighted dimensions.

4.1 Rubric Design

Our judge evaluates each edit against the following criteria:

(1) Instruction compliance (40% weight): Does the edit fulfill the specified changes? For addition tasks, the judge verifies that the target pathology is clearly visible and matches the instruction's description (e.g., location, extent, severity). For removal tasks, the judge checks that the image appears healthy and free of the target pathology. This dimension is weighted most heavily because instruction adherence is the primary goal.

(2) Structural plausibility (25% weight): Are anatomical structures reasonable and correct? The judge assesses whether rib cages, lung fields, brain ventricles, retinal vessels, etc., remain intact and properly aligned. It also checks for medical plausibility: do disease manifestations match known patterns (e.g., pneumothorax presenting as visceral pleural line separation, not random white patches)?

(3) Realism (20% weight): Does the image look like a real medical scan? The judge identifies visible artifacts such as blending inconsistencies, unnatural color shifts, repetitive structures (copy-paste patterns), or obvious deepfake signatures (e.g., checkerboard artifacts, local distortions).

(4) Fidelity preservation (15% weight): Does the image preserve the original imaging characteristics? This includes: (a) natural noise and grain texture appropriate to the modality; (b) device-specific artifacts like watermarks, grid lines, or scanner margins; (c) absence of sharp, unnatural edges around edited regions; (d) minimal intervention—only disease-relevant areas modified, all other regions unchanged. This dimension enforces our counterfactual minimal change principle.

Each dimension is evaluated as a boolean pass/fail, and the overall `qualified` verdict is true only if all four dimensions pass. The judge also provides a free-text `reason` explaining its decision, which we preserve for transparency and error analysis.

4.2 Judge Implementation

The judge receives three inputs: (1) the edited image, (2) the original image (for fidelity comparison), and (3) the instruction text. The system prompt (Appendix B) explicitly instructs the model to be thoughtful and balanced, accepting minor imperfections if the overall goal is achieved. This calibration reduces false negatives (rejecting acceptable edits due to nitpicking).

The judge outputs a JSON object structured as:

```
{
  "has_disease": true,          // instruction compliance
  "structure_reasonable": true, // structural plausibility
  "looks_realistic": true,      // realism
  "qualified": true,            // overall verdict
  "reason": "The edit successfully adds..."
}
```

Table 2: LLM-as-Judge pass rates by dimension, averaged across all editing attempts (including failed rounds). Common failure reasons are summarized from judge verdict texts.

Dimension	Pass rate (%)	Weight	Common failure reasons
Instruction compliance	78.3	40%	Pathology too subtle / overcorrection / wrong location
Structural plausibility	75.8	25%	Anatomical distortion / extra findings / non-target pathologies
Realism	79.1	20%	Visible blending artifacts / color inconsistencies / repetitive patterns
Fidelity preservation	71.2	15%	Unnatural edges / lost grain texture / excessive changes to non-target areas
Overall qualified	**66.0**	100%	Multiple dimension failures (typically fidelity + compliance)

4.3 Judge Performance Analysis

Table 2 reports dimension-level pass rates averaged across all editing attempts (both successful and failed rounds). Overall, 66.0% of editing attempts are qualified. The most common failure mode is **fidelity preservation** (71.2% pass rate), often triggered by unnatural edges or loss of grain texture. **Structural plausibility** (75.8% pass) fails when anatomical distortions or non-target pathologies appear. **Instruction compliance** (78.3% pass) and **realism** (79.1% pass) have higher pass rates, suggesting that the editing model generally succeeds at producing visually realistic outputs that approximately match instructions, but struggles with fine-grained medical constraints.

Inter-judge consistency. We randomly sample 500 edits and re-evaluate them with the same judge prompt but independent API calls (temperature=0.7). The judge produces identical verdicts for 87.2% of cases, and the remaining 12.8% differ by at most one dimension (e.g., realism passes in run A but fails in run B). This suggests reasonable but not perfect consistency, likely due to the stochastic nature of LLM reasoning. Future work could ensemble multiple judge runs or fine-tune a dedicated judge model for higher reliability.

Human validation (preliminary). We conduct a small pilot study with 2 board-certified radiologists, who independently review 100 chest X-ray edits (50 additions, 50 removals). Human verdicts agree with LLM-as-Judge in 81% of cases. Disagreements primarily occur when the judge is overly strict (rejecting acceptable edits, 12%) or overly lenient (accepting borderline cases, 7%). This preliminary result suggests that our LLM-as-Judge rubric provides a reasonable proxy for expert evaluation at scale, though systematic human validation remains a priority for future work.

5 Data Format, Organization, and Access

5.1 Directory Structure and File Naming

Successful edits are organized under `{dataset}-{task}` directories, with subdirectories per disease. Failed attempts are stored under `{dataset}-{task}-failed` with round-specific suffixes. All file names follow the pattern `{image_id}-{disease}-{task}[-failed-{round}].jpeg`. Example structure for MIMIC chest X-ray addition task:

```
MIMIC_single_disease_selection_dim1024_1k_per_class-edit/
  Pneumothorax/
    s50000230-Pneumothorax-edit.jpeg
    s50001145-Pneumothorax-edit.jpeg
    ...
  Atelectasis/
    s50002456-Atelectasis-edit.jpeg
    ...
  all_conversations.json    // centralized conversation logs
  final_prompts.json        // final instructions (success + failed)
  progress.json             // checkpoint for resumption
```

```
  api_failures.json         // API-level errors

MIMIC_...-edit-failed/
  Pneumothorax/
    s50002345-Pneumothorax-edit-failed-1.jpeg  // round 1 attempt
    s50002345-Pneumothorax-edit-failed-2.jpeg  // round 2 attempt
    ...
    s50002345-Pneumothorax-edit-failed-5.jpeg  // round 5 attempt
  failed_summary.json       // detailed failure summaries
```

This design ensures that successful edits are cleanly separated for downstream use, while preserving all intermediate attempts and metadata for research on refinement dynamics.

5.2 Metadata Schema

Each task entry in `all_conversations.json` is a dictionary with the following fields:

- `image_id`: unique identifier from source dataset (e.g., `s50000230`)
- `dataset`: source name (e.g., `MIMIC_single_disease_selection`)
- `modality`: one of `chest_xray`, `brain_mri`, `fundus`
- `disease`: pathology label (e.g., `Pneumothorax`, `Glioma`)
- `task`: `add` or `remove`
- `rounds`: list of dictionaries, each containing:
  - `round`: integer (1–5)
  - `instruction`: English editing prompt
  - `verdict`: JSON with `has_disease`, `structure_reasonable`, `looks_realistic`, `qualified`, `reason`
- `outcome`: `success` or `failed`
- `success_round`: integer (present only for successful tasks)
- `output_path`: relative path to the saved image

This schema enables straightforward loading into dataframes or databases for analysis. We provide Python example code for loading and parsing the metadata in the repository.

5.3 Licenses and Access

Source data. Original clinical images are subject to source-specific licenses and are *not* included in our release:

- **MIMIC-CXR:** PhysioNet Credentialed Health Data License (requires training and signed DUA)
- **Brain Tumor MRI:** Kaggle open dataset (CC0 or equivalent; verify at source)
- **ODIR-5K fundus:** Research use with attribution (specific terms at source)

Users must independently obtain access to source images if they wish to reproduce the editing pipeline or conduct comparative analysis.

Edited images and metadata. All edited images (both successful and failed), instructions, judge verdicts, and conversation logs are released under:

- **Images:** Creative Commons Attribution 4.0 International (CC BY 4.0)
- **Metadata (JSON, CSV):** Open Data Commons Attribution License (ODC-By 1.0)

Users are free to share, adapt, and build upon the data with proper attribution. We recommend citing this paper and the original source datasets.

Code and prompts. Construction scripts (`add_disease.py`, `remove_disease.py`, `retry_failed.py`, `make_assets.py`) and system prompts are released under the **MIT License** to facilitate reproducibility and extension.

Download. The dataset is hosted at `[URL placeholder]` with subsets available for each modality and task. Total size: approximately 140 GB compressed (190 GB uncompressed) for 88,457 1024×1024 JPEG images plus metadata. We provide SHA-256 checksums, a manifest file, and a brief usage tutorial in the repository README.

6 Ethics, Safety, and Limitations

6.1 Ethical Considerations

Patient privacy. All source datasets are de-identified according to their respective protocols and do not contain patient identifiers, clinical reports, or any information that could be traced to individuals. Our edited images are synthetic transformations of these de-identified sources and do not represent real patient conditions. No new private information is introduced during the editing process.

Intended use. Med-Banana-50K is designed *exclusively* for research purposes, including but not limited to: (1) training and evaluating medical image editing models, (2) counterfactual explanation and interpretability research, (3) data augmentation for diagnostic model training under controlled conditions, and (4) alignment and preference learning studies for multimodal LLMs. **The dataset is NOT intended for clinical diagnosis, treatment planning, or any direct patient care applications.** Users must not deploy models trained on this dataset in clinical settings without rigorous validation, regulatory approval (e.g., FDA clearance, CE marking), and continuous supervision by qualified healthcare professionals.

Misuse potential and mitigation. Synthetic medical image editing technology carries inherent risks of misuse. Potential harms include: (1) fabrication of medical records for insurance fraud or legal disputes; (2) generation of misleading educational materials that propagate incorrect medical knowledge; (3) creation of deceptive content to undermine trust in medical imaging; and (4) unauthorized alteration of patient records. We implement the following mitigations:

- **Transparent metadata:** All edited images are accompanied by JSON metadata documenting instructions, verdicts, and rounds, enabling provenance tracking.
- **Research-only licensing:** Our CC BY 4.0 license includes a strong ethical use recommendation, and we explicitly discourage clinical deployment without validation.
- **Watermarking (future):** We recommend that users deploying editing models embed invisible watermarks or cryptographic signatures to trace synthetic content.
- **Community guidelines:** We will maintain a dataset website with responsible use guidelines and a mechanism for reporting misuse.

6.2 Limitations and Future Work

Limited modality and disease coverage. Our dataset spans three modalities and approximately 20 disease types, representing a small fraction of clinical imaging. Generalization to other modalities (e.g., CT, ultrasound, histopathology, PET) and rare diseases remains untested. Future work should expand coverage to enable broader applicability.

LLM-as-Judge reliability. While our preliminary human validation study shows 81% agreement between LLM-as-Judge and radiologists, this is based on a small sample (100 edits, 2 radiologists). Systematic validation with larger expert panels, inter-rater reliability analysis, and calibration studies are essential to establish the rubric's trustworthiness. Additionally, LLM-as-Judge may inherit biases from its training data, potentially favoring certain visual styles or pathology presentations over others.

Editing model dependency. Our pipeline relies on Gemini-2.5-Flash-Image-Preview, a proprietary model whose capabilities, biases, and availability may change over time. Replication with open-source alternatives (e.g., InstructPix2Pix, LEDITS, BrushNet) would enhance reproducibility and community accessibility. Moreover, the editing model's training data and biases are not publicly documented, raising concerns about unknown failure modes or dataset contamination.

Quality variance. Despite rigorous quality control, some successful edits may still contain subtle artifacts, anatomical inconsistencies, or implausible disease presentations. Users should conduct domain-specific validation (e.g., radiologist review, downstream task performance) before deploying applications. We provide failed attempts and judge verdicts precisely to enable such critical analysis.

Counterfactual validity. While we enforce minimal change constraints, the clinical plausibility of synthetic pathologies has not been systematically validated by medical experts. Future work should include structured expert reviews to assess whether generated pathologies match real clinical presentations in terms of location, severity, co-occurrence patterns, and disease progression stages. Additionally, synthetic edits may not capture rare or atypical presentations, limiting their use for comprehensive augmentation.

Temporal and device variability. Our source images come from specific time periods and imaging devices with particular characteristics. Edits may not generalize to newer imaging protocols, different scanner manufacturers, or evolving clinical practices. Users should validate performance on their target datasets before deployment.

7 Related Work

General-domain image editing datasets. Recent years have seen the emergence of large-scale instruction-based editing datasets such as MagicBrush Zhang et al. [2024], InstructPix2Pix Brooks et al. [2023], and Pico-Banana-400K Qian et al. [2025]. These datasets focus on natural photography and generic editing operations (object addition/removal, style transfer, background changes). While valuable for general-purpose model training, they do not address medical domain constraints such as modality-specific noise, anatomical plausibility, or counterfactual minimal change.

Medical image synthesis and generation. Generative models have been applied to medical imaging for tasks including super-resolution Liu et al. [2021], modality translation Ye et al. [2017], and lesion synthesis Goodfellow et al. [2014]. However, these approaches typically focus on unconditional or class-conditional generation rather than instruction-based editing. Moreover, quality control often relies on FID or LPIPS metrics, which do not capture medical plausibility or instruction compliance.

Counterfactual medical imaging. Counterfactual generation has been explored for explainability and data augmentation Ghorbani et al. [2019], Singla et al. [2020]. These methods typically use GANs or diffusion models to generate "what-if" scenarios (e.g., removing a tumor to explain a diagnosis). Our work extends this paradigm by introducing instruction-based control, bidirectional editing, and scalable LLM-as-Judge quality assessment.

LLM-as-Judge for vision tasks. The use of LLMs to evaluate visual outputs has gained traction in text-to-image generation Fu et al. [2024], image captioning Zheng et al. [2023], and video quality assessment Ku et al. [2024]. Our medical rubric adapts this paradigm with domain-specific criteria (structural plausibility, fidelity preservation) and structured JSON outputs for programmatic analysis.

8 Conclusion

We present Med-Banana-50K, a cross-modality, bidirectional medical image editing dataset constructed through a systematic pipeline emphasizing instruction quality, single-step editing, LLM-as-Judge quality control, and history-aware refinement. With 50,635 successful edits and 37,822 failed attempts (88,457 total editing attempts) across chest X-ray, brain MRI, and fundus imaging, spanning 23 disease types, Med-Banana-50K provides a rich resource for training and evaluating medical image editing models, studying counterfactual generation, and exploring alignment and reflection in multimodal LLMs.

Our key contributions include: (1) a scalable construction pipeline enforcing medical fidelity, negative constraints, and minimal change; (2) a medically grounded LLM-as-Judge rubric with structured JSON verdicts and dimension-level analysis; (3) full release of conversations, prompts, and failed attempts to enable research on refinement dynamics and preference learning; and (4) comprehensive documentation, code, and metadata schemas to facilitate reproducibility and extension.

We acknowledge important limitations, including limited modality/disease coverage, LLM-as-Judge reliability requiring further human validation, and editing model dependency. Future work should

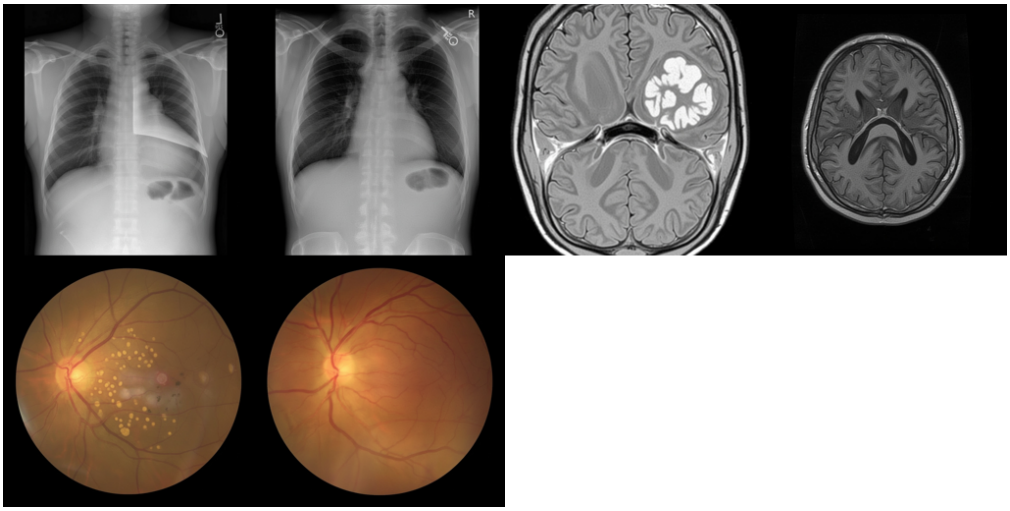

Figure 2: Representative edited results across modalities and tasks (add/remove).

expand to additional modalities, conduct systematic expert validation studies, replicate with open-source models, and explore applications in explainability, augmentation, and alignment research.

We hope Med-Banana-50K catalyzes progress in medical image editing and serves as a foundation for responsible, transparent, and medically grounded multimodal AI research. All data, code, and documentation are available at `[URL placeholder]` under permissive open licenses.

References

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jianquan Fu et al. GPT-4V(ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.15732*, 2024.

Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. HQ-Edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

Aaron Hurst et al. GPT-4o system card. OpenAI, 2024. URL `https://openai.com/research/gpt-4o-system-card`.

Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

Mingda Ku, Tianrui Li, Kai Zhang, Yan Lu, Xun Fu, and Weiping Zhu. Can large language models understand real-world complex instructions? *arXiv preprint arXiv:2309.09150*, 2024.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, et al. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021.

Masoud Nickparvar. Brain tumor MRI dataset. Kaggle, 2021. URL `https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset`.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-Banana-400K: A large-scale dataset for text-guided image editing. *arXiv preprint arXiv:2510.19808*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Peilun Shangguan, Jie Li, Wenhan Yang, et al. ODIR-5K: A large-scale Chinese ocular disease image database. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1664–1670, 2019.

Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. In *International Conference on Learning Representations (ICLR)*, 2020.

Menglong Ye, Edward Johns, Ankur Handa, Lourdes Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised Siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260*, 2017.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

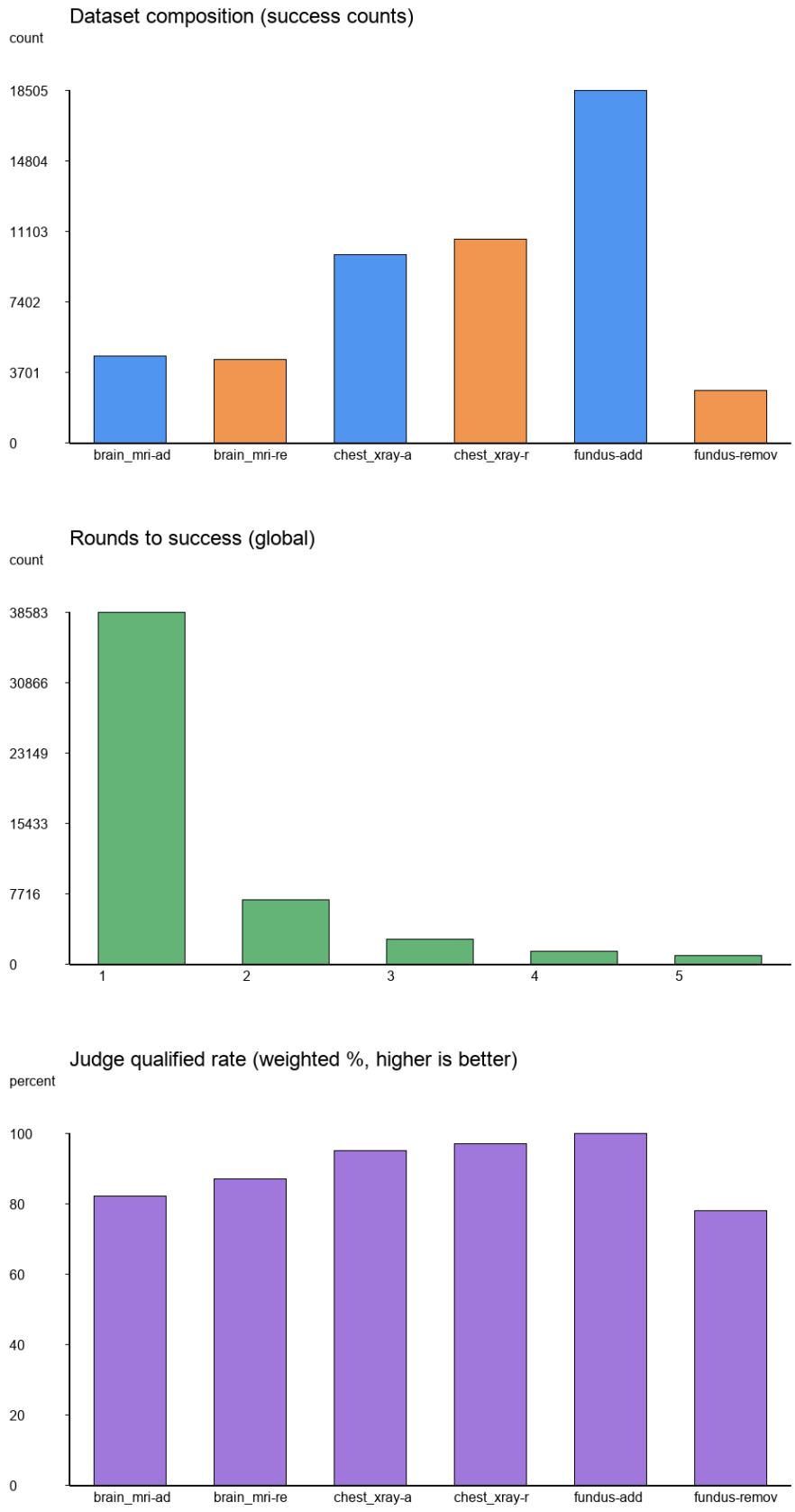

Figure 3: Top: dataset composition (success counts) by modality and task. Middle: rounds-to-success histogram. Bottom: judge qualified rates (weighted).