MIEDB-100K: A COMPREHENSIVE DATASET FOR MEDICAL IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

The scarcity of high-quality data remains a primary bottleneck in adapting multi-modal generative models for medical image editing. Existing medical image editing datasets often suffer from limited diversity, neglect of medical image understanding and inability to balance quality with scalability. To address these gaps, we propose MieDB-100k, a large-scale, high-quality and diverse dataset for text-guided medical image editing. It categorizes editing tasks into perspectives of Perception, Modification and Transformation, considering both understanding and generation abilities. We construct MieDB-100k via a data curation pipeline leveraging both modality-specific expert models and rule-based data synthetic methods, followed by rigorous manual inspection to ensure clinical fidelity. Extensive experiments demonstrate that model trained with MieDB-100k consistently outperforms both open-source and proprietary models while exhibiting strong generalization ability. We anticipate that this dataset will serve as a cornerstone for future advancements in specialized medical image editing. Dataset and code are publicly available at `https://anonymous.4open.science/r/MieDB-100k-3BB5`

1 INTRODUCTION

Multimodal generative models (Deng et al., 2025; Labs et al., 2025; Wu et al., 2025b; Liu et al., 2025c; DeepMind, 2025a; OpenAI, 2025; Li et al., 2026; Zhao et al., 2026) have developed rapidly in recent years. In natural image domains, generative models are not only gradually unifying text-guided generation and editing tasks, but also progressively expanding their capabilities to encompass image modification and image understanding (Deng et al., 2025; Tong et al., 2025; Li et al., 2026; Zhao et al., 2026). However, in medical image domains, their performance remains conspicuously limited, especially in the area of unified editing tasks (Liu et al., 2025b; Yang et al., 2025). We attribute this degradation primarily to a fundamental scarcity of specialized medical image-editing data.

While a few contemporary studies have proposed benchmarks or datasets for medical image editing, they remain insufficient in three key aspects: **(1) limited diversity in medical image modalities.** Unlike general computer vision, clinical imaging encompasses diverse modalities with distinct physical and structural foundations. However, existing research and datasets are restricted to a narrow range of imaging modalities (Chen & Feng, 2025; Liu et al., 2025b), typically the widely available modalities such as Chest X-rays and CTs, which cannot adequately train or evaluate a model's ability across diverse clinical settings.

(2) Neglect of medical image understanding. Almost all medical image editing works focus only on conceptual modification and stylistic transformation tasks, but ignore visual perception tasks (*e.g.*, organ/lesion detection), which has been considered to be beneficial to the generation of image editing models Huang et al. (2025); Deng et al. (2025); Tong et al.. This clinical grounding ensures interpretability and corrects 'right-for-the-wrong-reason' edits, which is vital for safety-critical medical applications. Moreover, recent work (Gabeur et al., 2026) has shown that generative models can develop strong visual understanding capabilities, suggesting that neglecting perception tasks may hinder the potential of medical generative models to bridge understanding and generation.

(3) Failure to ensure both data quality and scalability. Collection of medical image editing data is hindered by the difficulty of generating ground-truth counterfactuals. Some existing studies (Chen & Feng, 2025; Yang et al., 2025) distill general-purpose generative models for quick data scaling. However, these models are not tailored for medical use and hence produce results that lack clinical

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

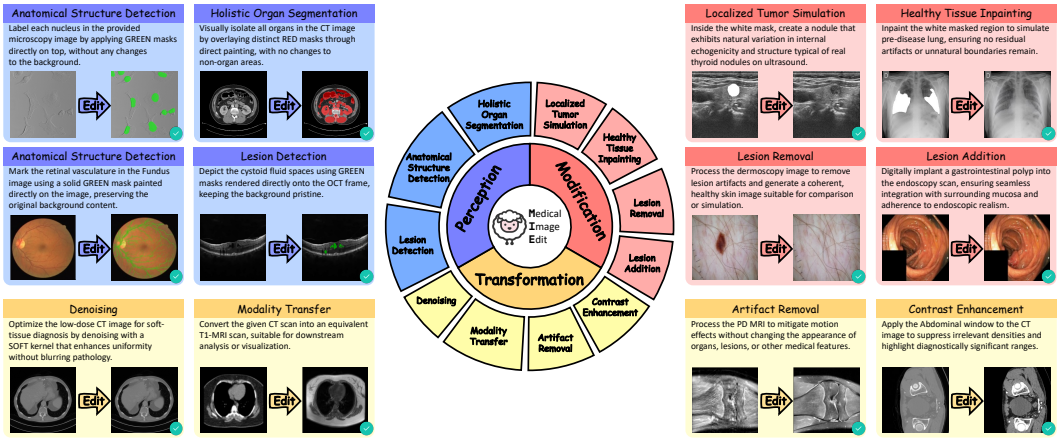


Figure 1: **MieDB-100k overview.** It categorizes medical image editing tasks into three perspectives, covering diverse medical modalities.

reliability and explainability. Conversely, previous work (Liu et al., 2025b) relies on extensive human involvement to manually collect real medical image pairs, which is notoriously difficult to scale up. Moreover, real-world longitudinal data often exhibits spatial misalignment and background inconsistency, as obtaining perfectly calibrated scan pairs is rare in practical medical settings.

In this paper, we address the aforementioned limitations in previous research by introducing **MieDB-100k**, a large-scale, high-quality, and diverse dataset for text-guided medical image editing. **MieDB-100k** includes **104,267** editing data, covering **63** distinct editing targets and **10** diverse medical image modalities. We categorize editing tasks into three types: **Perception**, **Modification** and **Transformation**, which consider both model's intrinsic understanding and generation abilities on medical images. To enhance the data fidelity while preserving the scalability, we propose a data curation pipeline leveraging both modality-specific expert models and rule-based data synthetic methods. Additionally, for some complex tasks such as lesion modification, we introduce individuals with medical knowledge to perform manual quality checks on the data to ensure data quality. Finally, we introduced task-specific evaluation metrics to facilitate a comprehensive assessment of the editing models' performance.

We evaluate existing open-source and closed-source multi-modal generative models on MieDB-100k and argue that most of them cannot perform well in medical image editing. To further validate the reliability and utility of MieDB-100k, we finetune the OmniGen2 baseline on our dataset. Experimental results demonstrate that MieDB-100k facilitates a substantial performance leap in medical image editing tasks, surpassing or matching SOTA models including Nano Banana Pro. It also exhibits strong generalization ability driven by the synergy of understanding and generation tasks. We anticipate that this dataset will serve as a cornerstone for future advancements in specialized medical image editing.

Our contributions can be summarized as follows: (1) We propose **MieDB-100k**, a large-scale, high-quality and highly diverse dataset for medical image editing with 63 targets and 10 medical image modalities. It is constructed via a credible and scalable data curation pipeline. (2) We first **unify the medical image understanding and generation** into the paradigm of edit, and find that joint training yields performance gains for specific tasks. (3) We evaluate popular open-source and closed-source multimodal generative models on **MieDB-100k**, and observe that training with our data can significantly strengthen the model's capacity for medical image editing.

2 Related Work

2.1 Data Research for Medical Image Editing

As an emerging area, multimodal medical generative modeling is currently supported by relatively few publicly available datasets for training and benchmarking (Tab. 1). In these works, the primary

challenge lies in the construction of high-quality image-edit pairs. MedEBench (Liu et al., 2025b), an early benchmarking effort, curated pairs by manually collecting related images from medical documents. While this ensures clinical validity, the approach lacks scalability. Furthermore, the resulting image pairs often exhibit background inconsistencies, as achieving strict spatial calibration in real-world clinical settings is virtually impossible. Conversely, Med-banana-50K (Chen & Feng, 2025) proposed a fully autonomous pipeline where data construction and quality control were managed by Gemini. However, applying general-purpose models to specialized medical scenarios may introduce factual errors or inconsistent edits, raising concerns about data fidelity. Finally, MedGEN-Bench (Yang et al., 2025) introduced image-edit pairs using a mix of rule-based and model-based methods; however, the lack of specific architectural details hinders a thorough evaluation of their data quality. Moreover, existing benchmarks only focus on content generation evaluation, overlooking the critical aspect of medical image understanding. To the best of our knowledge, MieDB-100k is the first work to incorporate medical image editing tasks spanning all three perspectives: Perception, Modification, and Transformation.

Table 1: **Comparison of contemporary medical image editing benchmarks and datasets. P** stands for Perception, **M** stands for Modification, and **T** stands for Transformation.

Benchmark	Size	Modalities	Targets	Perspectives	Source	Human Inspection
MedE-Bench (Liu et al., 2025b)	∼1k	4	13	M	Real	✓
Med-banana-50K (Chen & Feng, 2025)	∼50k	3	23	M	Synthetic	✗
MedGEN-Bench (Yang et al., 2025)	∼6k	6	16	M, T	Real & Synthetic	✓
MieDB-100k (Ours)	**∼100k**	**10**	**63**	**P, M, T**	**Real & Synthetic**	✓

2.2 Multimodal Generative Model

Multimodal generative models (Liu et al., 2025c; Sauer et al., 2024; Brooks et al., 2023) accept both images and natural language instructions as input, performing edits by translating semantic commands into precise visual manipulations. Recent studies (Wu et al., 2025a) often leverage vision-language model encoder and large-scale vision-language pretraining to align the semantic instruction with image modification. For instance, OmniGen2 (Wu et al., 2025b) utilizes Qwen2.5-VL (Bai et al., 2025b) to extract latent representations for semantic alignment, supported by a large-scale, multi-task training strategy. Furthermore, many recent studies (Deng et al., 2025; Li et al., 2026; Zhao et al., 2026) integrate image understanding and editing within a unified architecture. Exploiting these synergies(Tong et al.) is essential for creating robust models that are capable of performing both multimodal understanding and visual generation. On the commercial front, SOTA proprietary models like Gemini-3-Pro-Image (Nano Banana Pro) (DeepMind, 2025a) exhibit sophisticated image manipulation abilities, further realizing the real-world potential of multi-modal generative models. Despite these advancements, current models still struggle with the complexities of medical imaging(Liu et al., 2025b; Yang et al., 2025), highlighting the urgent need for comprehensive datasets to accelerate their adaptation to clinical domains.

3 MieDB-100k

This section introduces MieDB-100k, a high-quality, rigorous, and highly diverse dataset for medical image editing with more than 63 associated medical targets. It contains 104,267 image-editing triplets. Figure 2(a) summarizes the distribution of samples across 10 imaging modalities.

3.1 Data Definition

Each entry in MieDB-100k is a triplet (I, P, O), where I is the input medical image, P is the textual prompt that describes edit operation, and O is the target image.

3.2 Three Perspectives of MieDB-100k

MieDB-100k is constructed under a novel categorization of three perspectives, considering both understanding and generation capabilities: (1) **Perception** tasks, which focus on model's intrinsic medical knowledge via pixel-wise identification of prompted clinical targets in the input image;

(2) **Modification** tasks, which require the model to locate and alter specific medical features; and (3) **Transformation** tasks, involving medical image restoration, enhancement, and other low-level transformation. To ensure the rigor of the data triplets while maintaining scalability, we designed and implemented a specialized data construction pipeline for MieDB-100k (Fig 3), and we list all source datasets used for construction in App. A.

3.2.1 PERCEPTION

Perception tasks focus on medical image understanding, and we formulate it as an editing task by instructing the model to generate masks over regions of interest (ROIs), such as specific organs or lesions, through textual prompts. Notably, to align with image editing paradigm, the model is prompted to overlay the localization mask directly onto the source image rather than generating a standalone binary mask. This task serves two primary functions: First, since the mask-painting task only requires minimal pixel manipulation (typically modifying a single channel within a specific region), it serves as a **direct assessment of the medical knowledge embedded in the generative model**, isolating its perceptual accuracy from complex synthesis capabilities. Second, it introduces a promising application for multimodal generative models in the medical domain: **assisted interpretation in multimodal manner**. By allowing users to highlight specific targets in medical image through natural language prompts, this approach can assist patients in understanding their diagnostic images, aid medical students in their education, and reduce screening time for senior clinicians.

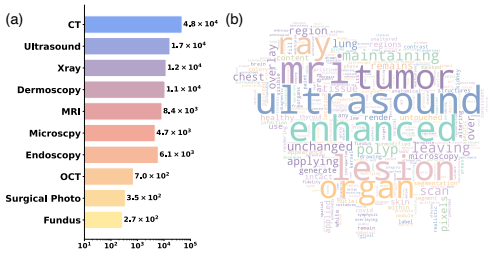


Figure 2: **Modality distribution (a) and prompt word cloud (b).**

The rule-based construction process for the Perception task's data triplets is illustrated in Fig. 3. Specifically, for a segmentation dataset, the original image serves as the input I. The output image O is synthesized by overlaying the ground-truth segmentation label, which is rendered in a randomly selected color (red, green, or blue), onto the input image with an alpha-blending transparency of 0.6. The ROIs of perception can be classified into three types: anatomical structure (organ, organism and so on), lesion area and holistic segmentation (segment all visible and clinically significant structures). We specify the perception target and visualizing color scheme in the textual prompt P. Since this part of the data is constructed following a definite rule, it can be readily scaled up to a diverse set of medical knowledge assessments and to the associated training dataset by leveraging the extensive body of existing medical segmentation research. Finally, to ensure a high-quality final benchmark, we manually filtered the initial data pool to remove trivial, redundant, or incorrectly labeled samples.

3.2.2 MODIFICATION

The perspective of Modification is specifically designed for semantically modifying medical contents, so as to address the diverse requirements of editing beyond just locating them. However, constructing modification data triplets is challenging because counterfactual image pairs cannot be captured simultaneously in the real world. While one could theoretically leverage general-purpose generative models (*e.g.*, Nano Banana Pro or Qwen-Image-Edit) to produce these edits, such models are not specialized for the medical domain, and therefore are prone to severe hallucinations, which is unacceptable in a healthcare context. To construct rigorous edit triplets and preserve scalability, we propose a four-stage process (Fig. 3) designed to bridge the gap between task complexity and model competence so as to fully utilize these automatic tools.

Stage I: We develop a suite of modality-specific expert models for healthy tissue inpainting, built upon the FLUX.1-Fill-dev model. This strategy is based on the observation that generating healthy anatomical structures is more stable and predictable than generating lesions, as the former exhibits more tractable patterns and textures. For each modality, we curate a training dataset consisting exclusively of non-pathological samples from existing medical image repositories. Through parameter-efficient finetuning, these models learn to inpaint masked areas with high clinical accuracy. We further apply background restoration and edge blending to correct any unintended modifications

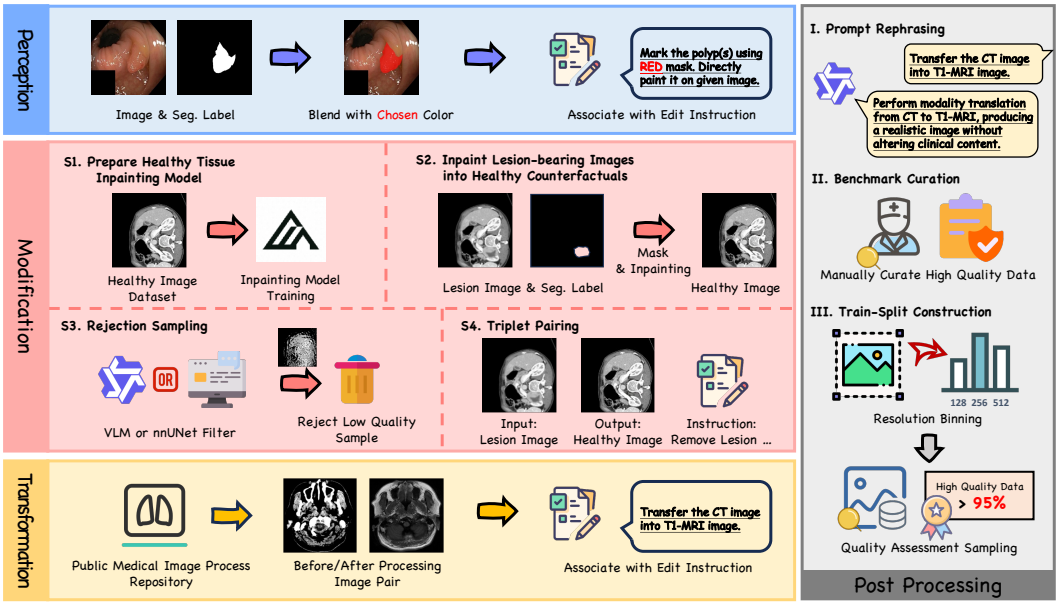


Figure 3: **Construction pipeline of MieDB-100k.**

made by the FLUX model outside the mask, ensuring the edited region blends seamlessly into the original image.

Stage II: We leverage these expert models to modify lesion-bearing images (L) into their counterfactual 'healthy' results (H). Specifically, we fill the lesion area in L using white pixels based on its ground-truth segmentation label. This masked image and its corresponding binary mask are then processed by the modality-matched expert model to synthesize H, where healthy tissue replaces lesion. Compared to distilling general-purpose generative models, our modality-specific approach not only restricts the high-variance generative process to a localized region to guarantee background consistency during the edit, but also ensures that tasks remain within the model's learned distribution, thereby significantly reducing hallucinations. Furthermore, unlike manual data collection from the internet, our approach provides superior scalability and efficiency.

Stage III: We implement a rejection sampling mechanism for the generated 'healthy' images (H) to further enhance the data quality within the Modification tasks. For modalities that resemble natural images (*e.g.*, endoscopy and dermoscopy), we prompt the Qwen3-VL-32B-Instruct model (Bai et al., 2025a) to filter out H that still contain lesions, exhibit artifacts, or are of low quality. For other modalities, we train separate nnUNet models (Isensee et al., 2021) for lesion segmentation and discard H where lesions remain detectable. For further details, please refer to App. B.3.

Stage IV: Triplet combination. Using these high quality 'lesion-healthy' counterfactual pairs, we generate diverse Modification task data by swapping the roles of L and H as input and output and varying the textual prompts P.

3.2.3 TRANSFORMATION

Transformation tasks include a wide array of low-level medical image processing operations. Unlike the localized edits found in Perception and Modification categories, tasks in this category typically require a holistic transformation of the entire input image.

The rule-based construction pipeline of Transformation tasks is shown in Fig. 3. From public repositories, we compile medical image pairs (I and O) representing 11 distinct transformation targets under four typical low-level vision categories. We then design specialized textual prompts P for each task to unify diverse medical image processing functions into the image editing framework.

3.2.4 POST PROCESSING

Prompt rephrasing. To enhance linguistic diversity, we utilize the Qwen-Max model to rephrase the prompts P for each data triplet. We also illustrate the linguistic diversity of our prompts via a word cloud in Fig. 2(b).

Benchmark curation. The training and test split of source dataset are strictly followed during the construction of MieDB-100k to preclude any data leakage. Furthermore, we recruit three people with clinical background to **manually evaluate and curate** 3,397 of the most representative samples characterized by high clinical fidelity from raw data test split (Please refer to App. B.4 for details) to serve as the benchmark of MieDB-100k, and we keep their original image size to minimize information loss.

Train split construction. For train split, we establish three resolution bins (128, 256, and 512) and resize images to their nearest corresponding value. To check the fidelity of training split, we randomly select 6,000 triplets for clinician evaluation, and **over 95% are viewed as high quality**.

3.3 MIEDB-100K EVALUATION

We evaluate MieDB-100k through two approaches: (1) verifiable metrics for the Perception and Transformation tasks, amenable to reward design in prevailing reinforcement learning (Shao et al., 2024; Liu et al., 2025a); and (2) more subjective evaluations for the Modification tasks, reflecting their greater complexity.

3.3.1 VERIFIABLE EVALUATION

Localization Accuracy Metric. We use the **DICE** Score for evaluating the spatial overlapping performance in Perception tasks. Notably, reconstructing a binary mask from the colored regions of an edited image is mathematically feasible when the background image and overlay color are known, and we detail this process in App. D.1, where the average DICE between original and reconstructed masks is 0.999 for grayscale images and 0.970 for RGB images. This procedure is applied to both model's output O_M and the ground truth images O to derive the mask of model's perceptual region and the ground truth region for DICE calculation.

To differentiate between models that accurately identify specific medical targets and those that merely generate coarse-grained masks, we further propose **Perception Accuracy**: The result is considered successful only if the DICE score exceeds a threshold of $\tau_{\text{DICE}} = 0.8$. This metric allows us to analyze whether a model possesses the specialized medical knowledge required for image understanding.

Image Similarity Metrics. We utilize **PSNR** and **SSIM** (Wang et al., 2004) to evaluate the similarity between the ground-truth and edited images at both the pixel and structural levels. For evaluations within the Perception perspective, we mask out the pixels corresponding to the ground-truth segmentation in both images. This allows us to specifically assess the model's ability to preserve the background while performing the requested edit. Moreover, to test the clinical utility of models in transformation perspective, we also exemplify a task-based evaluation in App. E.6.

3.3.2 EVALUATION FOR MODIFICATION TASKS

Vision-Language Model Rubric Scoring. Automating reliable assessments in the Modification tasks is inherently challenging, as edits are defined semantically and cannot be evaluated via deterministic rules. Existing benchmarks often leverage Vision-Language Models (VLMs) for this purpose, and we standardize the process and mitigate potential critic hallucinations by implementing a rubric-based scoring system. Specifically, we provide the VLM with the input image I, edit instruction P, reference output O, and the model's generated result O_M. Guided by the rubric, the VLM then performs a holistic evaluation of O_M.

We design a comprehensive scoring rubric (App. D.2) that assesses both the fulfillment of the editing intent and the model's ability to preserve the background. We utilize GPT-5.2 as an automated evaluator for this process, and map the final score to $[0, 100]$. We also test Gemini-3-flash as an alternative evaluator in App. D.2.3, where the Spearman rank correlation between two evaluators is **0.951**, demonstrating that our rubric-based protocol is robust to evaluator choice.

Table 2: **Overall result on MieDB-100k benchmark.** P-ACC means Perception Accuracy; B-PSNR and B-SSIM mean only calculate PSNR and SSIM on background pixels respectively; Rubric-S stands for the Rubric Score from VLM and Pref-Rank stands for human preference ranking. Best values are marked in red while second bests are in blue.

		Perception				Modification		Transformation	
Model Name	**Size**	**DICE**	**P-ACC**	**B-PSNR**	**B-SSIM**	**Rubric-S**	**Pref-Rank**	**PSNR**	**SSIM**
Open-Source									
SDXL-turbo (Sauer et al., 2024)	3.5B	0.002	0.000	16.6	0.467	8.4	7.7	15.4	0.369
Bagel (Deng et al., 2025)	7B	0.263	0.069	13.9	0.620	34.4	6.2	12.8	0.435
OmniGen2 (Wu et al., 2025b)	7B	0.248	0.065	11.9	0.541	29.1	7.1	9.1	0.305
Step1X-Edit (Liu et al., 2025c)	21B	0.332	0.126	15.5	0.727	35.6	4.5	16.0	0.533
Qwen-Image-Edit (Wu et al., 2025a)	27B	0.387	0.153	15.4	0.722	32.2	5.5	18.8	0.598
FLUX.1-Kontext-dev (Labs et al., 2025)	12B	0.347	0.126	15.4	0.701	37.8	6.2	17.3	0.514
OmniGen2-MIE (Ours)	7B	**0.831**	**0.737**	**28.1**	**0.917**	**65.9**	**1.4**	**22.2**	**0.668**
Closed-Source									
GPT-Image-1 (OpenAI, 2025)		0.467	0.221	16.3	0.510	42.8	4.8	14.2	0.430
Nano Banana Pro (DeepMind, 2025a)		0.426	0.202	12.8	0.413	63.4	2.0	19.8	0.639
Imagen4 (DeepMind, 2025b)		0.142	0.000	8.9	0.210	19.7	7.4	8.0	0.171

Human Preference Ranking. For each test case, we present the original triplet (I, P, O) and the outputs of all tested models simultaneously to evaluators, who are then asked to rank the various model-generated results according to their preference. By forcing this comparative ordering of all models, we are able to move beyond absolute quality scores and capture the relative strengths and weaknesses of current generative frameworks in a clinical setting. Specifically, we recruit 3 evaluators with clinical backgrounds to assess and rank the images edited by the benchmarked models, and compute the average ranking.

Evaluation Agreement. The Pairwise Spearman correlations among the three evaluators are 0.963, 0.989, and 0.957 (all $p < 1 \times 10^{-5}$), demonstrating good inter-rater reliability. Moreover, the correlation between human preference ranking and VLM rubric score is $\rho = -0.905$ ($p = 3 \times 10^{-4}$). Since smaller Pref-Rank means better, the strong negative correlation confirms that the VLM evaluator aligns well with human clinical judgment.

4 Experiments

4.1 Baselines

We evaluate nine models on MieDB-100k, comprising six open-source models: Qwen-Image-Edit-2511 (Wu et al., 2025a), Bagel (Deng et al., 2025), OmniGen2 (Wu et al., 2025b), Step1X-Edit-v1p2 (Liu et al., 2025c), FLUX.1-Kontext-dev (Labs et al., 2025) and SDXL-turbo (Sauer et al., 2024), plus three closed-source models: Nano Banana Pro DeepMind (2025a), GPT-Image-1 (OpenAI, 2025), and Imagen4 (DeepMind, 2025b). We implement open-source models following their official inference settings.

To validate the effectiveness of MieDB-100k, we finetune the OmniGen2 baseline on the training split and subject it to the same evaluation protocol as the other models. Specifically, we train the Diffusion Transformer (DiT) component for 20,000 iterations, employing a global batch size of 64 and a learning rate of 1e-4. We also finetune FLUX.1-Kontext baseline model to demonstrate the generality of our dataset across different model architectures. Please refer to App. E.3 for details.

4.2 Quantitative Results

We report the benchmarking results of MieDB-100k in Tab. 2. First, the extremely low **perception accuracy** indicates that **all tested models except ours** fail to accurately comprehend and localize the specified anatomical targets under our evaluation protocol. Consequently, in **Modification tasks**, most of them are unable to generate clinically meaningful edits. Although a few models, such as Nano Banana Pro, achieve competitive results, we are indeed observing the 'right-for-the-wrong-reason' phenomenon, a risk that must be strictly avoided in clinical settings. Since the poor performance in Perception tasks exposes their intrinsic lack of necessary medical knowledge, their edits cannot be justified. Notably in **Transformation tasks**, Nano Banana Pro also presents competitive results in certain cases. This may be attributed to the similarity between tasks like denoising or artifact removal

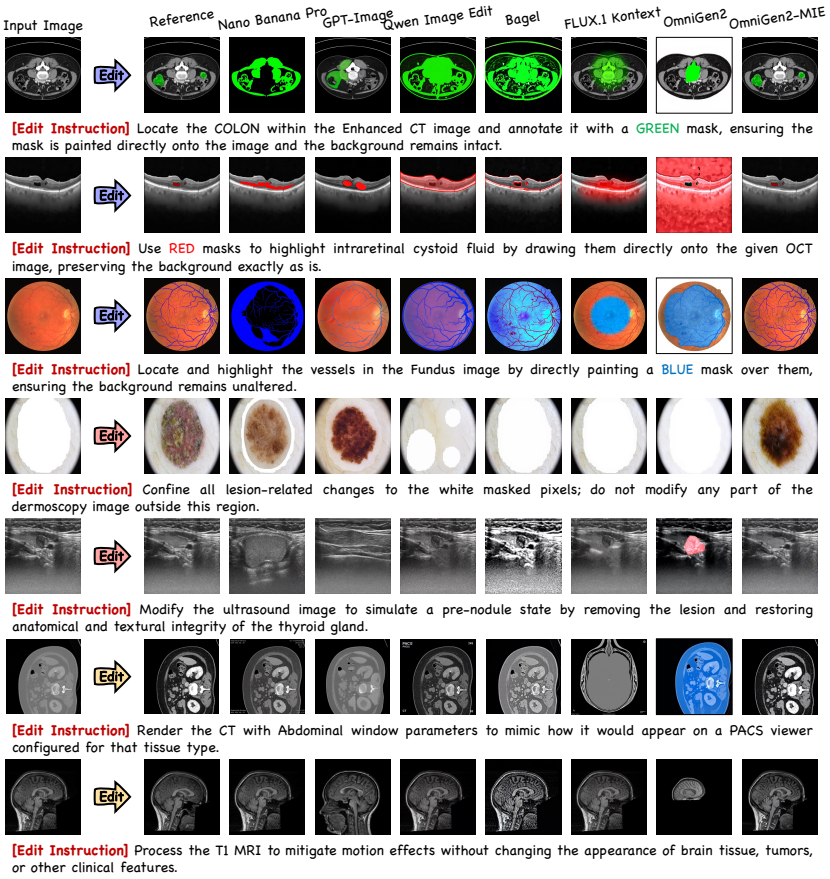


Figure 4: **Qualitative editing result comparison.**

and general-purpose low-level vision tasks, for which the model already possesses some capability (Zuo et al., 2025). Alternatively, it is possible that similar medical image processing tasks were included in its training set. Regardless, its absolute performance remains insufficient for practical clinical deployment. **In summary, the benchmark result demonstrates that current multimodal generative model cannot meet the requirement of medical imaging editing.**

Conversely, after training on MieDB-100k, a standard baseline model can achieve superior medical editing capabilities. As shown in Tab. 2, the OmniGen2-MIE model delivers the best performance across all three editing perspectives. The most significant improvements are observed in the Perception perspective, which demonstrate that MieDB-100k can effectively inject essential medical knowledge, thereby enhancing the interpretability of downstream editing tasks. Furthermore, in the Modification and Transformation tasks, where general-purpose editing abilities transfer more readily, our enhanced model still yields superior editing results compared to Nano Banana Pro. These findings highlight the pivotal role of our dataset in domain adaptation and establish a foundation for the development of understanding-generation unified medical models.

4.3 Qualitative Results

Fig. 4 presents qualitative editing results for several baseline models across the diverse modalities and tasks in MieDB-100k. These results demonstrate that the finetuned model exhibits an enhanced capability in both understanding and generation, allowing it to navigate the inherent complexities of medical image editing. Moreover, despite being explicitly prompted, even sophisticated closed-source models such as Nano Banana Pro fail to maintain background consistency in certain tasks. While their instruction-following proficiency stems from large-scale pre-training on natural image pairs, these capabilities tend to degrade when the distribution of medical modalities deviates significantly from

the natural images seen during pre-training. To further study the impact of modality deviation, we conduct a modality-wise analysis in App. E.1, and the results prove our judgment. This observation underscores the necessity of a highly diverse dataset like MieDB-100k to equip models with the capacity to handle a vast range of medical imaging modalities.

4.4 Ablation Study

To investigate the contribution of each task category, we conduct an ablation study by training models on individual perspective of MieDB-100k. We again utilize OmniGen2 as baseline model, following the training recipe described above while varying only the training data. As shown in Tab 3, each specialized model significantly outperforms the original baseline in its respective domain, validating the high information density and clinical relevance of our data.

Table 3: **Ablation study result on MieDB-100k.** **P** stands for Perception, **M** stands for Modification, and **T** stands for Transformation. Best values are marked in red, second bests are in blue.

	Perception		Modification	Transformation	
Training Data	**DICE**	**ACC**	**RubricScore**	**PSNR**	**SSIM**
Baseline (No train)	0.248	0.065	29.1	9.1	0.305
P-only	**0.833**	**0.740**	37.8	19.2	0.609
M-only	0.001	0.000	57.5	19.1	0.606
T-only	0.034	0.000	15.0	**23.9**	**0.696**
MieDB-100k	0.831	0.737	**65.9**	22.2	0.668

For the model trained on the full dataset, it achieves comparable or even better performance on all three perspectives, showing the effectiveness of the joint training (size-controlled ablation in App. E.4). More importantly, we observe significant performance improvement in the Modification perspective, demonstrating visual understanding ability has the potential to enhance visual generation ability. In summary, the ablation study shows that **MieDB-100k provides a synergistic training signal, enabling a versatile model for diverse medical editing tasks.**

4.5 Generalization Test

To investigate the cross-task synergy and the resulting generalization capabilities, we conduct an out-of-distribution editing experiment. We target 'bone metastasis', a medical target included in Perception tasks but strictly excluded from the Modification training data. We then prompt the OmniGen2-MIE model to perform metastasis addition and removal in CT scans.

As shown in Fig. 5, OmniGen2-MIE significantly outperforms OmniGen2 on this unseen task, demonstrating that our unified training on MieDB-100k can enhance the model's generalization capabilities across editing tasks. We also observe that Nano Banana Pro achieves the best OOD editing performance, marginally surpassing OmniGen2-MIE. We attribute this to its massive-scale general and medical editing data, underscoring the need to scale up such data. We also conduct an external evaluation on MedE-Bench (App. E.9) to corroborate the generalization ability.

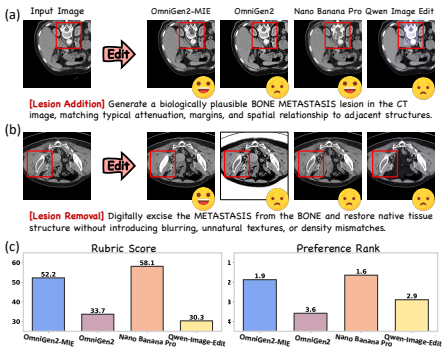


Figure 5: **Generalization test assessment.** (a) and (b): Edit samples output by different models on bone metastasis addition (a) and removal (b) tasks. Red bounding boxes are added post-hoc to highlight the edited regions for visualization; (c): Quantitative assessments following the recipe of Modification task evaluation.

5 Conclusion

In this paper, we introduce MieDB-100k, a large-scale and diverse dataset for text-guided medical image editing. By unifying Perception, Modification, and Transformation tasks into the paradigm of editing, our dataset bridges the gap between medical image understanding and generation. We develop a robust curation pipeline, integrating modality-specific expert models with rule-based synthesis, and enforce rigorous manual quality control to ensure clinical fidelity across all data. Extensive benchmarking demonstrates that model trained on MieDB-100k consistently outperforms both SOTA open-source and proprietary multimodal models while exhibiting exceptional generalization to unseen clinical tasks. Our work thus provides the data foundation to support the development and evaluation of multimodal generative models for clinical applications.

AI USE STATEMENT

LLMs/VLMs are used as core components in the data construction and evaluation pipeline: Qwen3-VL-32B-Instruct for rejection sampling in Modification tasks, Qwen-Max for prompt rephrasing, and GPT-5.2 as the automated evaluator for rubric scoring. All usages are explicitly described and reviewed. Additionally, we used generative AI tools for editing paper to improve readability. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The research uses only publicly available, de-identified medical image datasets. All source datasets have undergone their own institutional de-identification and ethical review processes. The paper discusses limitations and potential risks in Appendix H.

REPRODUCIBILITY STATEMENT

The data construction pipeline is fully described in Section 3.2. All training hyperparameters for OmniGen2-MIE are provided in Appendix C. Evaluation metrics and protocols are detailed in Section 3.3 and Appendix D. All 48+ source datasets are listed in Appendix A. The dataset and code are publicly released at `https://anonymous.4open.science/r/MieDB-100k-3BB5`.

REFERENCES

AAPM. Low dose ct grand challenge, 2016. URL `https://www.aapm.org/grandchallenge/lowdosect/#trainingData`. Accessed: 2026-09-24.

Amir Hossein Abdi, Shohreh Kasaei, and Mojdeh Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, 2(4):044003–044003, 2015.

Zeeshan Ahmed, Shahbaz Qamar Panhwar, Attiya Baqai, Fahim Aziz Umrani, Munawar Ahmed, and Arbaaz Khan. Deep learning based automated detection of intraretinal cystoid fluid. *International Journal of Imaging Systems and Technology*, 32(3):902–917, 2022.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Sharib Ali, Noha Ghatwary, Debesh Jha, Ece Isik-Polat, Gorkem Polat, Chen Yang, Wuyang Li, Adrian Galdran, Miguel-Ángel González Ballester, Vajira Thambawita, et al. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *Scientific Reports*, 14(1):2032, 2024.

Angiographics. Chuac dataset. URL `https://figshare.com/s/4d24cf3d14bc901a94bf`. Accessed: 2026-09-24.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Degerli Aysen, Chowdhury Muhammad, Rahman Tawsifur, Khandakar Amith, Qiblawey Yazan, Ahishali Mete, Tahir Anas M., and kiranyaz serkan. Qata-cov19 dataset, 2024. URL `https://www.kaggle.com/datasets/aysendegerli/qatacov19-dataset`. Accessed: 2026-09-24.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren,

Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *Preprint at arXiv*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *Preprint at arXiv*, 2025b. URL https://arxiv.org/abs/2502.13923.

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*, pp. 18392–18402, 2023.

Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109:218–225, 2019.

Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019.

Adrian Carballal, Francisco J Novoa, Carlos Fernandez-Lozano, Marcos García-Guimaraes, Guillermo Aldama-López, Ramón Calviño-Santos, José Manuel Vazquez-Rodriguez, and Alejandro Pazos. Automatic multiscale vascular image segmentation algorithm for coronary angiography. *Biomedical Signal Processing and Control*, 46:1–9, 2018.

Fernando Cervantes-Sanchez, Ivan Cruz-Aceves, Arturo Hernandez-Aguirre, Martha Alicia Hernandez-Gonzalez, and Sergio Eduardo Solorio-Meza. Automatic segmentation of coronary arteries in x-ray angiograms using multiscale analysis and artificial neural networks. *Applied Sciences*, 9(24):5507, 2019.

Zhihui Chen and Mengling Feng. Med-banana-50k: A cross-modality large-scale dataset for text-guided medical image editing. *Preprint at arXiv*, 2025. URL https://arxiv.org/abs/2511.00801.

Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8: 132665–132676, 2020.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, Ken Chang, Gennaro D'Anna, Lisa Deptula, Diviya Gupta, Muhammad Ammar Haider, Ali Hussain, Michael Iv, Marinos Kontzialis, Paul Manning, Farzan Moodi, Teresa Nunes, Aaron Simon, Nico Sollmann, David Vu, Maruf Adewole, Jake Albrecht, Udunna Anazodo, Rongrong Chai, Verena Chung, Shahriar Faghani, Keyvan Farahani, Anahita Fathi Kazerooni, Eugenio Iglesias, Florian Kofler, Hongwei Li, Marius George Linguraru, Bjoern Menze, Ahmed W. Moawad, Yury Velichko, Benedikt Wiestler, Talissa Altes, Patil Basavasagar, Martin Bendszus, Gianluca Brugnara, Jaeyoung Cho, Yaseen Dhemesh, Brandon K. K. Fields, Filip Garrett, Jaime Gass, Lubomir Hadjiiski, Jona Hattangadi-Gluth, Christopher Hess, Jessica L. Houk, Edvin Isufi, Lester J. Layfield, George Mastorakos, John Mongan, Pierre Nedelec, Uyen Nguyen, Sebastian Oliva, Matthew W. Pease, Aditya Rastogi, Jason Sinclair, Robert X. Smith, Leo P. Sugrue, Jonathan Thacker, Igor Vidic, Javier Villanueva-Meyer, Nathan S. White, Mariam Aboian, Gian Marco Conte, Anders Dale,

Mert R. Sabuncu, Tyler M. Seibert, Brent Weinberg, Aly Abayazeed, Raymond Huang, Sevcan Turk, Andreas M. Rauschecker, Nikdokht Farid, Philipp Vollmuth, Ayman Nada, Spyridon Bakas, Evan Calabrese, and Jeffrey D. Rudie. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *Preprint at arXiv*, 2024. URL https://arxiv.org/abs/2405.18368.

Google DeepMind. Gemini 3 pro: High-precision multimodal reasoning, 2025a. URL https://deepmind.google/models/gemini-image/pro/. Accessed: 2026-09-24.

Google DeepMind. Imagen 4.0: Model documentation and generation guide, 2025b. URL https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/imagen/4-0-generate. Accessed: 2026-09-24.

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. *Preprint at arXiv*, 2025. URL https://arxiv.org/abs/2505.14683.

Nicola Dietler, Matthias Minder, Vojislav Gligorovski, Augoustina Maria Economou, Denis Alain Henri Lucien Joly, Ahmad Sadeghi, Chun Hei Michael Chan, Mateusz Koziński, Martin Weigert, Anne-Florence Bitbol, et al. A convolutional neural network segments yeast microscopy images with high accuracy. *Nature communications*, 11(1):5723, 2020.

Viet Hung Duong, Huan Vu, Huong Duong Phan, Duc Quyen Nguyen, Duc Hao Pham, Quang Toan Le, Ba Sy Nguyen, Tien Dung Do, Viet Sang Dinh, Tien Cuong Nguyen, et al. Thyroidxl: Advancing thyroid nodule diagnosis with an expert-labeled, pathology-validated dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 616–626. Springer, 2025.

Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, Baiying Lei, Benjian Zhao, Xinxing Xu, Shaohua Li, Francisco Fumero, José Sigut, Haidar Almubarak, Yakoub Bazi, Yuanhao Guo, Yating Zhou, Ujjwal Baid, Shubham Innani, Tianjiao Guo, Jie Yang, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. *Preprint at arXiv*, 2022. URL https://arxiv.org/abs/2202.08994.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners. *Preprint at arXiv*, 2026. URL https://arxiv.org/abs/2604.20329.

Luis C Garcia-Peraza-Herrera, Lucas Fidon, Claudia D'Ettorre, Danail Stoyanov, Tom Vercauteren, and Sebastien Ourselin. Image compositing for segmentation of surgical tools without manual annotations. *IEEE transactions on medical imaging*, 40(5):1450–1460, 2021.

David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

Eri Haneda, Nils Peters, Jiayong Zhang, Grigorios Karageorgos, Wenjun Xia, Harald Paganetti, Ge Wang, Yi Guo, Jianhua Ma, Hyoung Suk Park, et al. Aapm ct metal artifact reduction grand challenge. *Medical physics*, 52(10):e70050, 2025.

Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, Taozhi Huang, Jiajia Liu, Qingpei Guo, Ming Yang, Jingdong Chen, and Jun Zhou. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *Preprint at arXiv*, 2025. URL https://arxiv.org/abs/2510.06590.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pp. 451–462. Springer, 2019.

Zeki Kuş and Musa Aydin. Medsegbench: A comprehensive benchmark for medical image segmentation in diverse data modalities. *Scientific Data*, 11(1):1283, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint at arXiv*, 2025. URL https://arxiv.org/abs/2506.15742.

Wenhui Lei, Hanyu Chen, Zitian Zhang, Luyang Luo, Qiong Xiao, Yannian Gu, Peng Gao, Yankai Jiang, Ci Wang, Guangtao Wu, Tongjia Xu, Yingjie Zhang, Pranav Rajpurkar, Xiaofan Zhang, Shaoting Zhang, and Zhenning Wang. A synthetic data-driven radiology foundation model for pan-tumor clinical diagnosis. *Preprint at arXiv*, 2025. URL https://arxiv.org/abs/2502.06171.

Lijiang Li, Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, and Chaoyou Fu. Omni-diffusion: Unified multimodal understanding and generation with masked discrete diffusion. *Preprint at arXiv*, 2026. URL https://arxiv.org/pdf/2603.06577.

Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint at arXiv*, 2025a. URL https://arxiv.org/abs/2505.05470.

Minghao Liu, Zhitao He, Zhiyuan Fan, Qingyun Wang, and Yi R Fung. Medebench: Diagnosing reliability in text-guided medical image editing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 767–791, 2025b.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *Preprint at arXiv*, 2025c. URL https://arxiv.org/abs/2504.17761.

Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637, 2012.

Yaosheng Lu, Mengqiang Zhou, Dengjiang Zhi, Minghong Zhou, Xiaosong Jiang, Ruiyu Qiu, Zhanhong Ou, Huijin Wang, Di Qiu, Mei Zhong, et al. The jnu-ifm dataset for segmenting pubic symphysis-fetal head. *Data in brief*, 41:107904, 2022.

Gendarme Mathieu, El Debs Bachir, et al. Brifiseg: a deep learning-based method for semantic and instance segmentation of nuclei in brightfield images. *Preprint at arXiv*, 2022. URL https://arxiv.org/abs/2211.03072.

Anna Montoya. Ultrasound nerve segmentation, 2026. URL https://www.kaggle.com/competitions/ultrasound-nerve-segmentation. Accessed: 2026-09-24.

Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzymyrskyy, Natalya V Ledikhova, Victor A Gombolevskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

Ádám Nárai, Petra Hermann, Tibor Auer, Péter Kemenczky, János Szalma, István Homolya, Eszter Somogyi, Pál Vakli, Béla Weiss, and Zoltán Vidnyánszky. Movement-related artefacts (mr-art) dataset of matched motion-corrupted and clean structural mri brain scans. *Scientific data*, 9(1):630, 2022.

Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38(2): 448–459, 2018.

Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In *International Symposium on Visual Computing*, pp. 15–28. Springer, 2021.

OpenAI. Gpt image 1 model documentation, 2025. URL https://platform.openai.com/docs/models/gpt-image-1. Accessed: 2026-09-24.

OpenAI. Gpt image 2 model, 2026. URL https://openai.com/index/introducing-chatgpt-images-2-0. Accessed: 2026-09-24.

Nikhil Pandey. Chest xray masks and labels, 2025. URL https://www.kaggle.com/datasets/nikhilpandey360/chest-xray-masks-and-labels. Accessed: 2026-09-24.

Marco Polo. Chest ct segmentation, 2025. URL https://www.kaggle.com/datasets/polomarco/chest-ct-segmentation. Accessed: 2026-09-24.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint at arXiv*, 2024. URL https://arxiv.org/abs/2402.03300.

Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022.

Christoph Spahn, Estibaliz Gómez-de Mariscal, Romain F Laine, Pedro M Pereira, Lucas von Chamier, Mia Conduit, Mariana G Pinho, Guillaume Jacquemet, Séamus Holden, Mike Heilemann, et al. Deepbacs for multi-task bacterial image analysis using open-source deep learning approaches. *Communications Biology*, 5(1):688, 2022.

Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in biology and medicine*, 139:105002, 2021.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, Naila Murray, Marjan Ghazvininejad, Mike Lewis, Nicolas Ballas, Amir Bar, Michael Rabbat, Jakob Verbeek, Luke Zettlemoyer, Koustuv Sinha, Yann LeCun, and Saining Xie. Beyond language modeling: An exploration of multimodal pretraining. *Preprint at arXiv*. URL https://arxiv.org/abs/2603.03276.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17001–17012, 2025.

David A Van Valen, Takamasa Kudo, Keara M Lane, Derek N Macklin, Nicolas T Quach, Mialy M DeFelice, Inbal Maayan, Yu Tanouchi, Euan A Ashley, and Markus W Covert. Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS computational biology*, 12(11):e1005177, 2016.

Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E Ahmed Raza, Nasir Rajpoot, et al. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12):3413–3423, 2021.

Vision and Image Processing Lab. Skin cancer detection, 2024. URL https://vip.uwaterloo.ca/skin-cancer-detection/. Accessed: 2026-09-24.

Santiago Vitale, José Ignacio Orlando, Emmanuel Iarussi, and Ignacio Larrabide. Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. *International journal of computer assisted radiology and surgery*, 15(2):183–192, 2020.

Yuting Wang and Feng Shi. KMAR-50K. Mendeley Data, V6, 2025. URL https://doi.org/10.17632/xw7mrg7ntg.6. Accessed: 2026-09-24.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5 (5):e230024, 2023.

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. *Preprint at arXiv*, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *Preprint at arXiv*, 2025b. URL https://arxiv.org/abs/2506.18871.

Junjie Yang, Yuhao Yan, Gang Wu, Yuxuan Wang, Ruoyu Liang, Xinjie Jiang, Xiang Wan, Fenglei Fan, Yongquan Zhang, Feiwei Qin, and Changmiao Wang. Medgen-bench: Contextually entangled benchmark for open-ended multimodal medical generation. *Preprint at arXiv*, 2025. URL https://arxiv.org/abs/2511.13135.

Linfeng Yang, Rajarshi P Ghosh, J Matthew Franklin, Simon Chen, Chenyu You, Raja R Narayan, Marc L Melcher, and Jan T Liphardt. Nuset: A deep learning tool for reliably separating and analyzing crowded cells. *PLoS computational biology*, 16(9):e1008193, 2020.

Zakaria Shams Siam Younus Akon. Paired ct and mri dataset for medical applications, 2025. URL https://www.kaggle.com/datasets/29c3607295965ebb030f2d158fec487412d84c82528dd44f8ef956aef35541aa. Accessed: 2026-09-24.

Siqi Zhang, Qizhe Zhang, Shanghang Zhang, Xiaohong Liu, Jingkun Yue, Ming Lu, Huihuan Xu, Jiaxin Yao, Xiaobao Wei, Jiajun Cao, et al. A generalist foundation model and database for open-world medical image segmentation. *Nature Biomedical Engineering*, pp. 1–16, 2025.

Yaqi Zhao, Wang Lin, Zijian Zhang, Miles Yang, Jingyuan Chen, Wentao Zhang, Zhao Zhong, and Liefeng Bo. Unicom: Unified multimodal modeling via compressed continuous semantic representations. *Preprint at arXiv*, 2026. URL https://arxiv.org/abs/2603.10702.

Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018.

Jialong Zuo, Haoyou Deng, Hanyu Zhou, Jiaxin Zhu, Yicheng Zhang, Yiwei Zhang, Yongxin Yan, Kaixing Huang, Weisen Chen, Yongtai Deng, Rui Jin, Nong Sang, and Changxin Gao. Is nano banana pro a low-level vision all-rounder? a comprehensive evaluation on 14 tasks and 40 datasets. *Preprint at arXiv*, 2025. URL https://arxiv.org/abs/2512.15110.

CONTENTS

A Data Sources **18**

B Construction Details **19**

B.1 Processing Flowchart 19

B.2 Inpainting Model Training Details 19

B.3 Automatic Filtering Details 19

B.3.1 nnUNet Filtering 19

B.3.2 Qwen3-VL-32B-Instruct Filtering 20

B.4 Manual Inspection Details 21

B.5 Preservation Rates 21

B.6 Target Distribution 22

C Implementation Details of OmniGen2-MIE **22**

D Evaluation Details **23**

D.1 Mask Reconstruction via Alpha De-blending 23

D.1.1 Mathematics 23

D.1.2 Case of mask reconstruction 23

D.2 VLM Automatic Scoring 23

D.2.1 VLM Scoring Rubric 23

D.2.2 Case 25

D.2.3 Alternative VLM Consistency 25

E Supplementary Experiments **26**

E.1 Modality-Wise Performance Analysis 26

E.2 Multi-Round Generation 26

E.3 Baseline Model Generalization 27

E.4 Size-Controlled Ablation 27

E.5 Out-Of-Distribution Image Edit 27

E.6 Transformation Downstream Validation 27

E.7 Evaluation of GPT-Image-2 28

E.8 Perception Accuracy Threshold Sensitivity Analysis 28

E.9 External Validation on MedE-Bench 29

E.10 Inpainting Quality Verification 29

E.10.1 Turing Test 29

E.10.2 Examples of Healthy Tissue Inpainting 29

F More Qualitative Results **31**

G Failure Cases **34**

H Limitations and Broader Impacts **34**

A Data Sources

Our work is compiled based on following public medical image repositories and inherits their demographic and institutional characteristics:

Table 4: **Summary of public medical datasets utilized in the construction of MieDB-100k.** The columns #Train and #Benchmark denote the number of samples allocated to our training and benchmark splits respectively from each source dataset.

DatasetName	#Train	#Benchmark	Modality	License
AbdomenUS (Vitale et al., 2020)	569	62	Ultrasound	CC BY 4.0
Bbbc010 (Ljosa et al., 2012)	70	20	Microscopy	CC BY 4.0
Bkai-Igh (Ngoc Lan et al., 2021)	700	81	Endoscopy	CC BY 4.0
Brats-gli (de Verdier et al., 2024)	1529	80	MRI	CC0
BriFiSeg (Mathieu et al., 2022)	1005	40	Microscopy	CC BY 4.0
BUSI (Al-Dhabyani et al., 2020)	452	80	Ultrasound	CC BY 4.0
CellNuclei (Caicedo et al., 2019)	469	51	Microscopy	CC BY 4.0
ChaseDB1 (Carballal et al., 2018)	19	7	Fundus	CC BY 4.0
Chest-ct-segmentation (Polo, 2025)	278	19	CT	CC0
Chest-xray-masks-and-labels (Pandey, 2025)	666	32	Xray	CC0
CHUAC (Angiographics)	17	5	Fundus	CC BY-NC 4.0
COVID-19_Radiography_Dataset (Chowdhury et al., 2020)	2010	95	Xray	MIT
COVID-19-CT-SCAN-Lesion (Morozov et al., 2020)	255	15	CT	CC BY 4.0
CovidQU (Tahir et al., 2021)	5684	122	Xray	CC BY-SA 4.0
CT_MAR (Haneda et al., 2025)	1595	82	CT	BSD 3-Clause
CT-Low-Dose-Reconstruction (AAPM, 2016)	867	51	CT	CC BY 4.0
CystoidFluid (Ahmed et al., 2022)	703	59	OCT	CC BY 4.0
Dca1 (Cervantes-Sanchez et al., 2019)	93	28	Fundus	CC BY-NC 4.0
Deepbacs (Spahn et al., 2022)	17	10	Microscopy	CC BY 4.0
Drive (Staal et al., 2004)	18	20	Fundus	CC BY 4.0
DynamicNuclear (Van Valen et al., 2016)	50	17	Microscopy	CC BY 4.0
FHPsAOP (Lu et al., 2022)	2800	80	Ultrasound	CC BY 4.0
IDRiD (Porwal et al., 2018)	47	27	Fundus	CC BY 4.0
ISIC2016 (Gutman et al., 2016)	810	80	Dermoscopy	CC0
ISIC2018 (Codella et al., 2019)	9973	115	Dermoscopy	CC0
KMAR-50K (Wang & Shi, 2025)	651	47	MRI	CC BY 4.0
Kvasir (Jha et al., 2019)	4429	139	Endoscopy	Custom (Research Only)
Lgg-mri-segmentation (Buda et al., 2019)	1669	55	MRI	CC BY-NC-SA 4.0
MoNuSAC (Verma et al., 2021)	0	21	Microscopy	CC BY 4.0
MR-ART (Nárai et al., 2022)	820	18	MRI	CC BY-NC 4.0
MSD (Antonelli et al., 2022)	797	39	MRI	Apache-2.0
NuSeT (Yang et al., 2020)	2383	40	Microscopy	CC BY 4.0
Paired_MRI_CT (Younus Akon, 2025)	1974	72	CT, MRI	Custom (Research Only)
Pandental (Abdi et al., 2015)	81	24	Xray	CC BY 4.0
Pasta-GEN (Lei et al., 2025)	32299	731	CT	MIT
PolypGen (Ali et al., 2024)	984	75	Endoscopy	CC BY 4.0
PROMISE12 (Litjens et al., 2014)	1031	80	MRI	CC BY 4.0
QaTa-COV19 (Aysen et al., 2024)	3573	85	Xray	CC BY-NC-SA 3.0
Refuge (Fang et al., 2022)	80	80	Fundus	CC BY 4.0
RoboTool (Garcia-Peraza-Herrera et al., 2021)	350	76	Surgical Photo	CC BY 4.0
ThyroidXL (Duong et al., 2025)	7029	138	Ultrasound	Custom (Research Only)
Tnbcnuclei (Naylor et al., 2018)	35	10	Microscopy	CC BY 4.0
TotalSegmentator (Wasserthal et al., 2023)	5206	154	CT, MRI	CC BY 4.0
UltrasoundNerve (Montoya, 2026)	1651	50	Ultrasound	CC BY 4.0
USforKidney (Song et al., 2022)	4351	50	Ultrasound	CC BY 4.0
UWSkinCancer (Vision & Lab, 2024)	143	44	Dermoscopy	CC BY 4.0
WBC (Zheng et al., 2018)	280	40	Microscopy	CC BY 4.0
YeaZ (Dietler et al., 2020)	358	51	Microscopy	CC BY 4.0

We also appreciate MedSegBench(Kuş & Aydin, 2024) and MedSegDB(Zhang et al., 2025) for collecting and pre-processing some of these datasets.

B CONSTRUCTION DETAILS

B.1 PROCESSING FLOWCHART

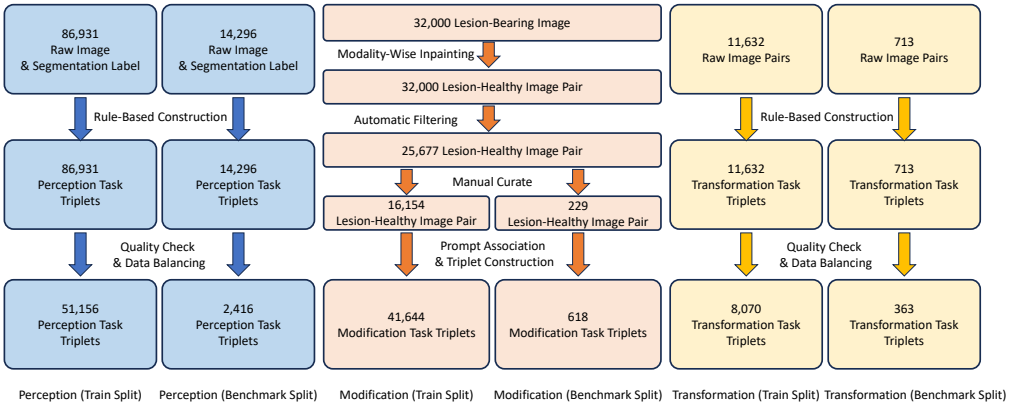


Figure 6: **Construction details of the three perspectives.** For Modification, all candidate pairs that pass automatic filtering undergo expert review (Good/Fair/Poor) and *Poor* pairs are discarded, while the benchmark split is curated from *Good* samples to uphold high clinical standards. Across the full training set, a separate sampling-based audit further confirms a high-quality proportion exceeding 95%.

B.2 INPAINTING MODEL TRAINING DETAILS

We use LoRA finetuning to train FLUX.1-Fill-dev. Implementation details of each modality experts are listed in Tab. 5. We conduct a simple hyper-parameter search over learning rate and number of training steps, selecting the final model based on both empirical inpainting performance and automatic rejection rate.

Hyper-Parameters	**CT**	**Ultrasound**	**MRI**	**Xray**	**Dermoscopy**	**Endoscopy**
Finetuning method	LoRA	LoRA	LoRA	LoRA	LoRA	LoRA
Rank	64	64	64	64	64	64
Steps	5,000	2,000	5,000	2,000	5,000	2,000
#GPUs	4	4	4	4	4	4
Per-device batch size	4	4	4	4	4	4
Global batch size (effective)	16	16	16	16	16	16
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-4}
Precision	BF16	BF16	BF16	BF16	BF16	BF16

Table 5: **Training hyper-parameters used for finetuning FLUX.1-Fill-dev into modality specific health-tissue inpainting expert.**

B.3 AUTOMATIC FILTERING DETAILS

B.3.1 NNUNET FILTERING

We use the default standard 5-fold cross-validation for nnUNet training following its official implementations. The model structure and training hyper-parameters are generated automatically via `nnUNetv2_plan_and_preprocess` tool. To prevent possible data leakage, we preclude the nnUNet 5-fold cross-validation training data from lesion-bearing images to be inpainted in Modification Stage II. The model is trained to generate segmentation map for given medical images. During rejection sampling, the sample is discarded if nnUNet still detects lesion-like pixels in the inpainted region.

B.3.2 QWEN3-VL-32B-INSTRUCT FILTERING

For each generated sample, we compose a three-panel collage consisting of: (1) the **inpainting mask** (left), which highlights the edited region in white; (2) the **edited image** H (center), where the lesion has been replaced with synthetic healthy tissue; and (3) the **original image** L (right), serving as a reference. To further aid the model's assessment, the edited region in the center frame is also highlighted with a red bounding box. Then, the collaged image, together with a modality-specific structured prompt, is fed to Qwen3-VL-32B-Instruct. The prompt instructs the model to evaluate the edited region on multiple clinically relevant dimensions and return a structured JSON response.

Based on the model output, we retain only the high-quality samples. Specifically, a generated image H is **accepted** only if its image quality is rated as "good" and no residual lesion or unnatural content is detected (rated as "no"). All other samples are **discarded**. This conservative filtering strategy ensures that only clinically plausible inpainting results enter the subsequent triplet combination stage.

Below we showcase the VLM prompt for Endoscopy filtering:

VLM Filter Prompt for Endoscopy

```
You are a specialized Medical AI Evaluator for endoscopic image
inpainting. Your task is to assess the quality of an AI-edited
image where a polyp has been removed and replaced with synthetic
tissue.

### Task
Evaluate the center frame (Edited Image) by comparing it with
the right frame (Original Image) and the left frame (Inpainting
Mask). You must determine if the polyp removal is medically
plausible or if the AI has introduced "unnatural content."
For clarity, the polyp area in the center frame has also been
highlighted by red bounding box.

### Focus Area
Concentrate your evaluation ONLY on the region indicated by the
WHITE mask in the leftmost frame.

### Criteria Definitions
- **image_quality**:
    - "good": Sharp mucosal texture, proper lighting, clear
      vascular patterns.
    - "borderline": Mild blur, slight overexposure/underexposure,
      but still interpretable.
    - "poor": Severe motion blur, pixelation, or lighting that
      obscures anatomy.
- **has_polyp_or_unnatural_content**:
    - "obvious": Presence of residual polyp, or medically
      impossible structures like an abrupt hole (perforation-like),
      a "crater," or a "donut" protrusion.
    - "minor": Unnatural flatness or slight bulge that doesn't
      look like healthy mucosa.
    - "no": The area looks like smooth, healthy, and continuous
      intestinal mucosa.

### Strict Output Format
Return ONLY a JSON object. Do not include any conversational
filler.
{
  "image_quality": "good" | "borderline" | "poor",
  "has_polyp_or_unnatural_content": "obvious" | "minor" | "no",
  "reasoning": "A concise explanation (max 2 sentences) focusing
      on why the content is considered unnatural or successful."
}
```

```
You are not making a medical diagnosis, only screening image
quality and obvious abnormalities.
Return ONLY a JSON object, no additional text or explanation.
```

B.4 Manual Inspection Details

We engage **three board-certified medical experts** to assess the images in MieDB-100k. The evaluation protocol differs by task type, reflecting the distinct nature of each data source:

Perception and Transformation tasks (rule-based data). We sample approximately 3,000 cases from the source test split and remove those with mask mismatches or transformation errors. Owing to the high quality of the source datasets, only a negligible number of samples are excluded. The remaining data constitute the MieDB-100k benchmark split for these tasks.

Modification tasks (model-generated data). All samples that passed the rejection-sampling stage are submitted to expert review to ensure clinical validity. The experts are informed of the intended use of the data (i.e., serving as medical image editing pairs) and are asked to classify each sample into three categories: **Good**. High-quality edits that successfully accomplish the intended modification without noticeable artifacts; **Fair**. Edits that achieve the intended modification with only slight, clinically negligible traces that do not affect their usability as medical editing data; **Poor**. Failed edits or samples with obvious artifacts or incorrect modifications.

Note that to support consistent evaluation, we present each case as a collage of the original image and the inpainted image, and provide a detailed instruction document including task-specific descriptions and representative examples.

Quality-check results. Final labels are determined by **majority vote** among the three experts, and more than 76% of the reviewed samples receive unanimous 3-out-of-3 agreement, indicating strong inter-rater consistency. The results are approximately evenly distributed across the three categories. We remove all *poor* samples and construct the benchmark split using only *good* samples. Concretely, of the 25,677 candidate pairs that passed automatic rejection sampling, the experts discarded 9,294 pairs rated as *poor*; the retained 16,383 *good/fair* pairs constitute the Modification training (16,154 pairs) and benchmark (229 pairs) splits. In addition, as a separate post-hoc audit, we randomly sample 6,000 cases from the full 100k training set (spanning all three perspectives) for spot-checking, and find that over 95% are of high quality.

B.5 Preservation Rates

Phase	CT	Ultrasound	MRI	Xray	Dermoscopy	Endoscopy
Automatic Filtering	86.1%	76.1%	61.6%	88.2%	64.9%	67.6%
Manual Inspection	64.0%	64.5%	49.9%	55.3%	73.2%	66.0%

Table 6: **Data preservation rates during automatic filtering and manual inspection.**

B.6 TARGET DISTRIBUTION

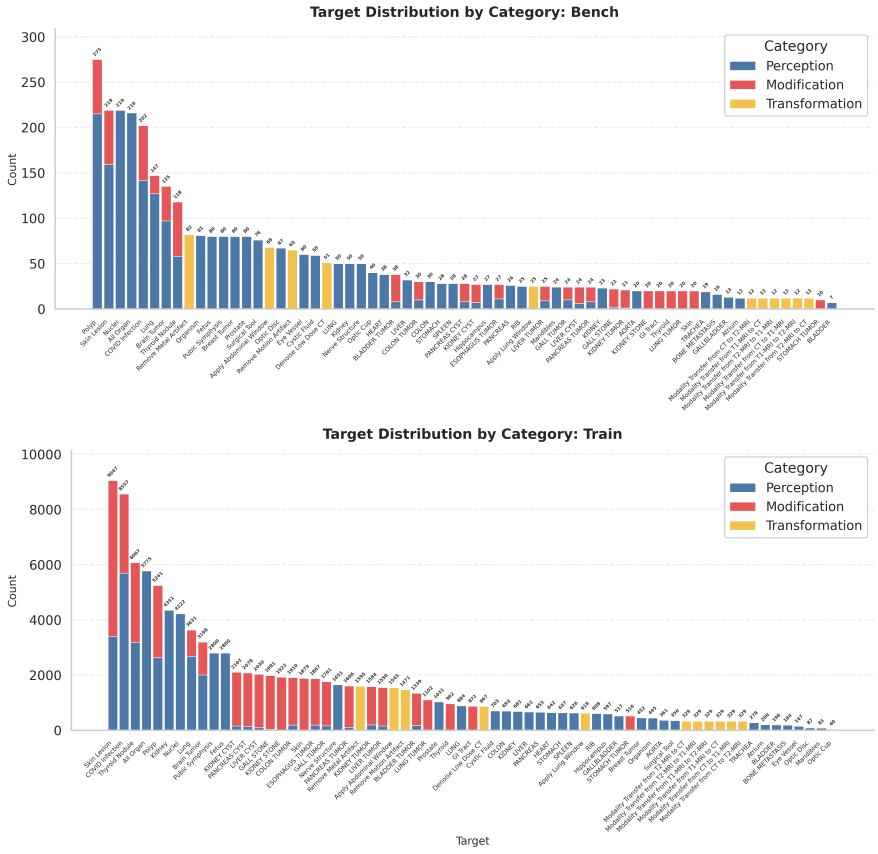


Figure 7: **Distribution details of edit target.**

C IMPLEMENTATION DETAILS OF OMNIGEN2-MIE

Hyper-Parameter	**Value**
Finetuning method	Full-Parameter Finetuning (DiT component)
snr_type	lognorm
do_shift	True
dynamic_time_shift	True
Steps	20, 000
#GPUs	8 × H20 (96G)
Per-device batch size	8
Gradient accumulation	1
Global batch size (effective)	64
Learning rate	1×10^{-4}
LR scheduler	timm_constant_with_warmup
Warm-up_t	500
Precision	BF16
Random seed	2233

Table 7: **Training hyper-parameters used for finetuning OmniGen2-MIE on our dataset.**

D Evaluation Details

D.1 Mask Reconstruction via Alpha De-blending

D.1.1 Mathematics

To recover the segmentation mask from the visualized output, we model the edited image $\mathbf{O}$ as a linear interpolation between the original background image $\mathbf{B}$ (a.k.a. the input image I) and a known overlay color $\mathbf{C}$ (red, green or blue). This relationship is governed by the per-pixel alpha channel $\alpha \in [0, 1]$, according to the standard alpha blending equation:

$$\mathbf{O} = (1 - \alpha)\mathbf{B} + \alpha\mathbf{C} \tag{1}$$

By rearranging the terms as $\mathbf{O} - \mathbf{B} = \alpha(\mathbf{C} - \mathbf{B})$, the scalar value α can be interpreted as the projection of the observed color shift onto the vector representing the maximum possible color change. To account for potential noise in the RGB space, we solve for α at each pixel using the least-squares solution:

$$\alpha = \frac{(\mathbf{O} - \mathbf{B}) \cdot (\mathbf{C} - \mathbf{B})}{|\mathbf{C} - \mathbf{B}|^2 + \epsilon} \tag{2}$$

For numerical stability, a small constant $\epsilon = 10^{-6}$ is added to the denominator to prevent division by zero in regions where the background color $\mathbf{B}$ approaches the overlay color $\mathbf{C}$. The continuous alpha map is subsequently binarized to produce the final segmentation mask M. This is achieved by applying a global threshold τ_α, such that:

$$M_{i,j} = \begin{cases} 1 & \text{if } \alpha_{i,j} > \tau_\alpha \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In our implementation, a threshold of $\tau_\alpha = 0.5$ is utilized to effectively separate the predicted regions from the background. The average DICE between original and reconstructed masks is 0.999 for grayscale images (e.g., X-ray) and 0.970 for RGB images (e.g., Endoscopy), confirming robustness to compression and color drift.

D.1.2 Case of mask reconstruction

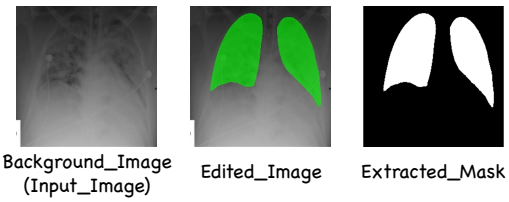


Figure 8: **Case of perception mask reconstruction.**

D.2 VLM Automatic Scoring

D.2.1 VLM Scoring Rubric

Scoring Rubric for Modification Tasks

You are a helpful assistant in evaluating medical image editing result.
You will be provided with an edit instruction and a collage image where the leftmost is origin image, center is edited image and rightmost is the reference ground truth image.
You should score how well an edited image matches the intended edit while preserving clinical realism and image integrity based on following scoring rubrics:

1) **Edit Goal Fulfillment (Edit Correctness)**: Assesses whether the intended lesion change is achieved.

Scoring reference:

- *5: Lesion added/removed exactly as intended; no residuals or unintended remnants.*
- *4: Mostly correct; slight residual signal after removal or slight under/over-addition.*
- *3: Partial success; lesion still partially present (removal) or incomplete/incorrect lesion (addition).*
- *2: Wrong area or wrong type of change; target lesion largely unchanged.*
- *1: No effective edit or opposite edit performed.*

2) **Edit Area Morphology (Shape, Margins, Internal Structure)**: Evaluates whether edit area matches expected morphology and/or reference.

Scoring reference:

- *5: Shape, border characteristics, and internal texture are highly consistent.*
- *4: Minor border/shape irregularities; still plausible.*
- *3: Morphology is generic/unconvincing; borders/texture inconsistent.*
- *2: Clearly artificial morphology (blocky, repeated patterns, unnatural contours).*
- *1: Morphology nonsensical or misleading (e.g., appears like different pathology).*

3) **Intensity / Signal / Attenuation Consistency**: Checks whether edited region match modality-specific intensities.

Scoring reference:

- *5: Intensities match local tissue statistics; no intensity discontinuities.*
- *4: Slight intensity mismatch detectable with careful viewing.*
- *3: Obvious mismatch (too bright/dark), inconsistent with modality or anatomy.*
- *2: Strong intensity discontinuity; clearly edited.*
- *1: Severe intensity errors that invalidate the image (e.g., saturation/clipping, inverted contrast).*

4) **Boundary Blending & Transition Naturalness**: Rates edge blending and transitions between edited and unedited regions.

Scoring reference:

- *5: Seamless blending; no halos, ringing, cut-paste edges.*
- *4: Minor halo/edge artifacts only on close inspection.*
- *3: Visible seams; boundary looks edited.*
- *2: Strong cutout appearance or blur patches.*
- *1: Boundary artifacts dominate the image.*

5) **Background / Non-target Preservation**: Measures unintended changes outside the lesion edit region. Scoring reference:

- *5: Non-target anatomy and background unchanged (within expected noise).*
- *4: Small unintended changes but not clinically meaningful.*
- *3: Noticeable unintended alterations in nearby structures.*
- *2: Large unintended modifications to anatomy or overall image.*
- *1: Global corruption or major anatomical distortions.*

6) **Anatomical Plausibility & Clinical Coherence**: Assesses whether result respects anatomy and pathology logic (e.g., lesion doesn't cross impossible boundaries). Scoring reference:

- *5: Fully plausible; consistent with organ boundaries and expected presentation.*
- *4: Mostly plausible; minor oddity but acceptable.*
- *3: Questionable plausibility (e.g., lesion overlaps structures unnaturally).*
- *2: Clearly implausible anatomy/pathology relationship.*
- *1: Clinically nonsensical or misleading.*

7) **Artifact Introduction (Noise, Texture, Aliasing, Compression, Repetition)**: Evaluates new artifacts introduced by editing. Scoring reference:
   - *5: No new artifacts; noise texture consistent with original.*
   - *4: Minor artifacts (subtle smoothing/grain mismatch).*
   - *3: Artifacts visible and distracting.*
   - *2: Strong artifacts (banding, checkerboard, repeated texture).*
   - *1: Severe artifacts preventing clinical use.*

8) **Image Quality & Acquisition Consistency**: Checks consistency with scanner characteristics (resolution, blur, point spread, slice thickness cues, motion). Scoring reference:
   - *5: Matches acquisition characteristics; sharpness/noise consistent.*
   - *4: Slight mismatch in sharpness or noise level.*
   - *3: Clear mismatch (over-smoothed or over-sharpened region).*
   - *2: Strong mismatch; edited region appears from different source.*
   - *1: Completely inconsistent with acquisition; unusable.*

Return Format: Return a JSON dictionary with two fields:

- "`conclusion`": A brief conclusion to the edited image.
- "`score_list`": The scores of the eight aspects in a JSON list.

D.2.2 CASE

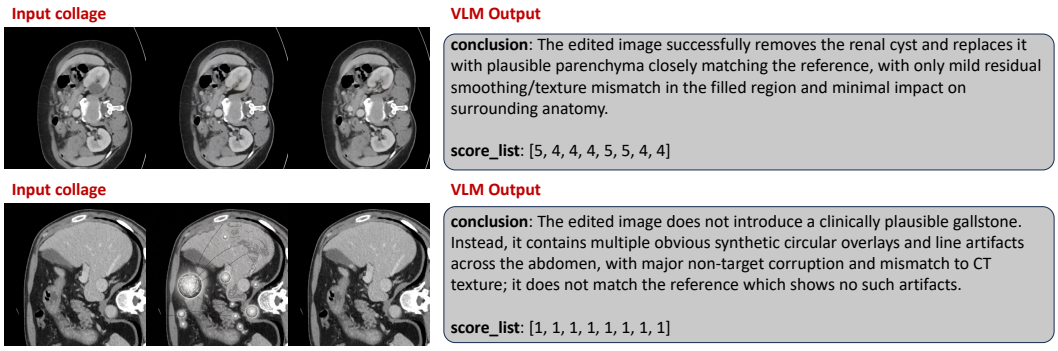


Figure 9: **Cases of VLM rubric scoring.**

D.2.3 ALTERNATIVE VLM CONSISTENCY

To validate the stability and robustness of our VLM rubric scoring methodology, we evaluate the Modification task outputs using two different VLM evaluators: Gemini-3-flash and GPT-5.2. As shown in Tab. 8, despite the absolute scores varying between evaluators, the relative ranking of all benchmarked models remains highly consistent. The Spearman rank correlation between the two evaluators is $\rho = 0.951$ ($p = 2 \times 10^{-5}$), indicating strong agreement. This cross-evaluator consistency demonstrates that our rubric-based evaluation protocol is robust to the choice of VLM backbone and reliably reflects the relative quality of different models' editing outputs.

Table 8: **VLM Rubric Score using different evaluators.**

Evaluator	OmniGen2-MIE	Gemini	GPT-Image	FLUX	Step1x	Bagel	Qwen-Image	OmniGen2	Imagen4	SDXL
Gemini-3-flash	37.0	35.9	26.1	18.6	21.0	19.8	10.8	13.4	8.9	7.3
GPT-5.2	65.9	63.4	42.8	37.8	35.6	34.4	32.2	29.1	19.7	8.4

E SUPPLEMENTARY EXPERIMENTS

E.1 MODALITY-WISE PERFORMANCE ANALYSIS

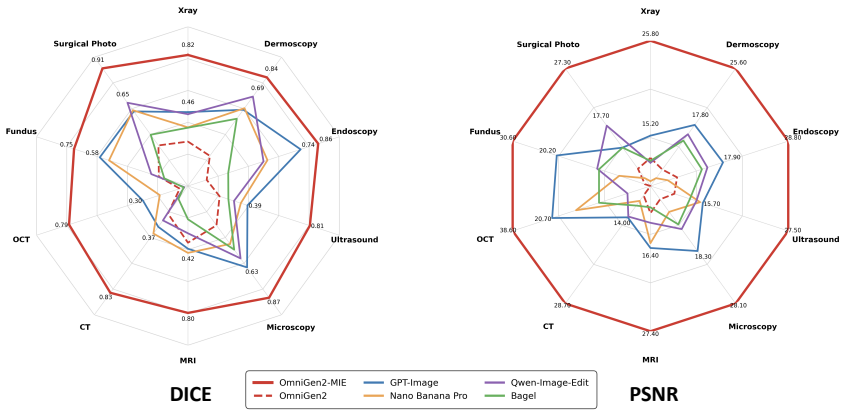


Figure 10: **Modality-wise performance analysis result within perception perspective.** Left: DICE score; right: PSNR score.

To investigate the impact of modality deviation, we conduct a modality-wise analysis of the benchmarking results within the Perception perspective. Specifically, we report the DICE and PSNR scores of six representative models across all medical imaging modalities included in MieDB-100k. As illustrated in Fig. 10, the experimental results are consistent with our hypotheses. For the baseline models, performance is unevenly distributed across the various modalities: They achieve relatively strong results on modalities that resemble natural images, such as Endoscopy, Dermoscopy, and Surgical Photo. However, on non-optical modalities (e.g., CT, MRI, Ultrasound), their performance degrades drastically. In contrast, the model trained on our dataset exhibits balanced and superior performance across all imaging types. Collectively, these results demonstrate that a diverse dataset like MieDB-100k is essential for successfully adapting multi-modal generative models to the medical domain.

E.2 MULTI-ROUND GENERATION

Table 9: **Multi-round generation result.** Best values are marked in **Bold**

	Perception								Transformation			
	DICE		P-ACC		B-PSNR		B-SSIM		PSNR		SSIM	
	Pass@1	Pass@3	Pass@1	Pass@3	Pass@1	Pass@3	Pass@1	Pass@3	Pass@1	Pass@3	Pass@1	Pass@3
Open-Source												
SDXL-turbo	0.002	0.003	0.000	0.000	16.6	17.0	0.467	0.484	15.4	15.7	0.369	0.396
Bagel	0.263	0.383	0.069	0.137	13.9	16.1	0.620	0.703	12.8	15.2	0.435	0.533
OmniGen2	0.248	0.357	0.065	0.125	11.9	14.4	0.541	0.628	9.1	15.3	0.305	0.528
Step1X-Edit	0.332	0.369	0.126	0.143	15.5	16.4	0.727	0.748	16.0	16.5	0.533	0.548
FLUX.1-Kontext-dev	0.347	0.41	0.126	0.174	15.4	16.5	0.701	0.761	17.3	18.9	0.514	0.578
Qwen-Image-Edit	0.387	0.493	0.153	0.249	15.4	17.4	0.722	0.795	18.8	20.2	0.598	0.645
OmniGen2-MIE (Ours)	**0.831**	**0.856**	**0.737**	**0.789**	**28.1**	**28.8**	**0.917**	**0.921**	**22.2**	**23.0**	**0.668**	**0.697**

To mitigate the inherent variance of the generative process, we report Pass@3 scores for the open-source models on Perception tasks. Specifically, we generate three independent outputs for each editing task and select the highest-performing sample to represent the task's score. These results are then averaged across all tasks to provide a robust assessment of overall performance.

The results of the multi-round generation tests are summarized in Tab. 9. While multi-round generation improves the absolute scores for baseline models, it does not alter the underlying fact that these models lack essential medical knowledge. Furthermore, the significant fluctuations across rounds expose the high-variance nature of these baselines, undermining their reliability under clinical applications. In contrast, our model exhibits remarkable stability across all three trials. This consistency suggests that model trained on MieDB-100k has developed a deterministic understanding of medical concepts rather than relying on fortuitous generation.

E.3 Baseline Model Generalization

Besides OmniGen2, we also finetune FLUX.1-Kontext(Labs et al., 2025) baseline model on MieDB-100k to demonstrate the generality of our dataset across different model architectures. With global batch of 64, we iterate the backbone transformer for 40,000 steps under learning rate of 5×10^{-6} using AdamW optimizer. As shown in Tab. 10, the finetuned FluxKontext-MIE model also achieves substantial improvements over the original FluxKontext baseline across all three editing perspectives. These results confirm that the benefits of training on MieDB-100k are not limited to a single architecture, but generalize across different multimodal generative models.

Table 10: **Result of finetuning FLUX.1-Kontext on MieDB-100k.** Best values are marked in red.

	Perception				Modification	Transformation	
Model	**DICE**	**P-ACC**	**B-PSNR**	**B-SSIM**	**Rubric-S**	**PSNR**	**SSIM**
FluxKontext(Labs et al., 2025)	0.347	0.126	15.4	0.701	37.8	17.3	0.514
FluxKontext-MIE	**0.627**	**0.468**	**25.7**	**0.916**	**51.3**	**23.1**	**0.726**

E.4 Size-Controlled Ablation

The ablation in Tab. 3 compares each single-perspective subset against the full mixture, which inevitably confounds cross-task synergy with the larger total training volume. To disentangle the two, we additionally train three models on **equal-sized** (40,000-triplet) subsets—**P-only**, **M-only**, and a **P+M** mixture—keeping the total data volume strictly constant.

Table 11: **Size-controlled ablation.** All rows use an identical total of 40,000 training triplets. P-DICE / P-ACC denote Perception DICE and Perception Accuracy; M-RubricScore denotes the Modification rubric score. Best values in **bold**.

Training Data	#P-Triplets	#M-Triplets	P-DICE	P-ACC	M-RubricScore
P-only	40,000	0	**0.825**	**0.734**	35.1
M-only	0	40,000	0.002	0.000	55.9
P+M	20,000	20,000	0.798	0.695	**61.2**

With the total volume held constant, the P+M model attains a Modification RubricScore of 61.2, surpassing M-only (55.9) by $+5.3$ while preserving strong Perception ability. This confirms that the Modification gain stems from perceptual grounding rather than a mere increase in data scale, corroborating the synergy hypothesis in Sec. 4.

E.5 Out-Of-Distribution Image Edit

While Section 4.5 demonstrates that the model trained on MieDB-100k generalizes effectively to OOD editing targets, we further evaluate its robustness by performing edits on 'in-the-wild' medical images sourced from the internet (Fig. 11).

The results indicate that our model is capable of readily adapting to medical images outside of datasets. This suggests that the diversity of MieDB-100k has successfully decoupled the model from specific data distribution, allowing it to internalize generalizable edit operations that are applicable to real-world clinical scenarios.

E.6 Transformation Downstream Validation

To further demonstrate the practical clinical value of MieDB-100k, we conduct a downstream validation experiment on the metal artifact removal (MAR) task from the Transformation perspective. Specifically, we apply a pretrained all-organ segmentation model to both the original artifact-corrupted CT images and the restored CT images produced by editing models. By comparing the segmentation accuracy on the restored images against the ground-truth masks, we can quantitatively assess whether the editing model preserves clinically relevant anatomical structures after artifact removal.

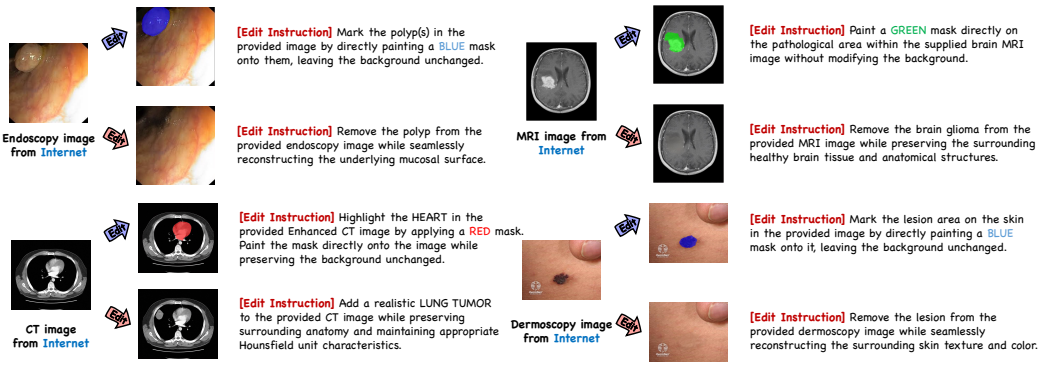


Figure 11: **Examples of Out-Of-Distribution Editing.**

Table 12: **Downstream segmentation validation on metal artifact removal.** DICE scores between ground-truth masks and segmentation results on restored CT images.

Restoration Model	Qwen-Image-Edit	Nano Banana Pro	OmniGen2	OmniGen2-MIE
DICE	0.472	0.699	0.355	**0.873**

As shown in Tab. 12, OmniGen2-MIE significantly outperforms all baseline models. This result demonstrates that the model trained on MieDB-100k not only produces visually plausible restorations, but also faithfully preserves the underlying anatomical structures that are critical for downstream clinical analysis. This experiment further validates that MieDB-100k possesses significant real-world relevance and clinical value: training on our dataset enables editing models to produce outputs that are genuinely useful for practical medical workflows such as computer-aided diagnosis on processed images.

E.7 Evaluation of GPT-Image-2

GPT-Image-2 (OpenAI, 2026) is a proprietary editing model released after our main benchmark batch (Tab. 2), and is therefore evaluated separately here.

Table 13: **Comparison with GPT-Image-2 on the MieDB-100k benchmark.** B-PSNR / B-SSIM are computed on background pixels for Perception; T-PSNR / T-SSIM denote Transformation. Best values in **bold**.

Model	**DICE**	**P-ACC**	**B-PSNR**	**B-SSIM**	**Rubric-S**	**T-PSNR**	**T-SSIM**
GPT-Image-2	0.581	0.407	22.5	0.841	62.8	16.7	0.515
OmniGen2-MIE (Ours)	**0.831**	**0.737**	**28.1**	**0.917**	**65.9**	**22.2**	**0.668**

GPT-Image-2 is a strong proprietary baseline (Perception DICE 0.581, higher than GPT-Image-1's 0.467 in Tab. 2), yet OmniGen2-MIE surpasses it across all three perspectives. This confirms that domain-specific training on MieDB-100k provides value that general-purpose pre-training alone cannot match, even against the latest proprietary systems.

E.8 Perception Accuracy Threshold Sensitivity Analysis

In Tab. 14, we present the result under different perception accuracy thresholds τ_{DICE} in Perception tasks evaluation. The conclusions are fully robust: OmniGen2-MIE maintains a dominant lead across all thresholds, while all baselines remain far below clinically useful levels regardless of threshold choice.

Table 14: Perception tasks performance comparison across various alternative thresholds.

Model	DICE	P-ACC ($\tau_{\text{DICE}} = 0.70$)	P-ACC ($\tau_{\text{DICE}} = 0.80$)	P-ACC ($\tau_{\text{DICE}} = 0.85$)	P-ACC ($\tau_{\text{DICE}} = 0.90$)
Qwen-Image-Edit	0.387	0.237	0.153	0.111	0.072
Nano Banana Pro	0.426	0.401	0.202	0.136	0.102
GPT-Image-2	0.581	0.497	0.407	0.326	0.251
OmniGen2	0.248	0.097	0.065	0.053	0.034
OmniGen2-MIE	**0.831**	**0.844**	**0.737**	**0.659**	**0.520**

E.9 External Validation on MedE-Bench

To provide generalization evidence beyond the single out-of-distribution target in Sec. 4.5, we further evaluate on **MedE-Bench** (Liu et al., 2025b), an independently curated benchmark built from real clinical image pairs with **no overlap** with our training sources.

Table 15: **External validation on MedE-Bench.** EA: Editing Accuracy; CP: Contextual Preservation; VQ: Visual Quality (VLM-based metrics computed with the official code). All metrics are higher-is-better.

Model	CLIP	PSNR	SSIM	EA	CP	VQ
OmniGen2	0.234	14.6	0.898	0.35	0.64	0.51
OmniGen2-MIE (Ours)	**0.267**	**15.4**	**0.951**	**0.46**	**0.77**	**0.57**

Training on MieDB-100k yields consistent improvements across **all six metrics** under an independent, real-reference evaluation protocol, confirming that the capabilities acquired from MieDB-100k transfer effectively to clinical evaluation settings beyond our own benchmark.

E.10 Inpainting Quality Verification

E.10.1 Turing Test

To directly verify the quality of our FLUX-based healthy tissue inpainting, we conduct a Turing Test. Specifically, we collage 100 real healthy images and FLUX-generated healthy images side by side (with their left-right order randomized) and ask 3 medical experts to judge whether the image on the left is Good (more realistic), Same (comparable in quality), or Bad (more likely to be generated) relative to the image on the right.

Table 16: **Turing Test results on FLUX-generated healthy images.** Good indicates the generated image fooled the expert; Same indicates comparable quality; Bad indicates the expert correctly identified the generated image.

Rating	Expert 1	Expert 2	Expert 3
Good (gen fooled)	4%	8%	9%
Same (equal quality)	86%	77%	83%
Bad (gen identified)	10%	15%	8%

As shown in Tab. 16, the vast majority of generated images were rated as Same or Good by all three experts, with the "Same" category consistently at least 77%. This directly confirms that the inpainted healthy tissue achieves quality comparable to real medical images under expert clinical judgment, validating the fidelity of our Modification task data construction pipeline.

E.10.2 Examples of Healthy Tissue Inpainting

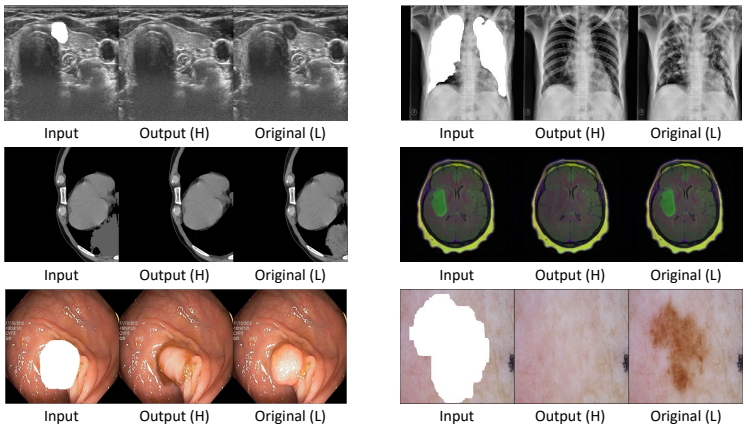
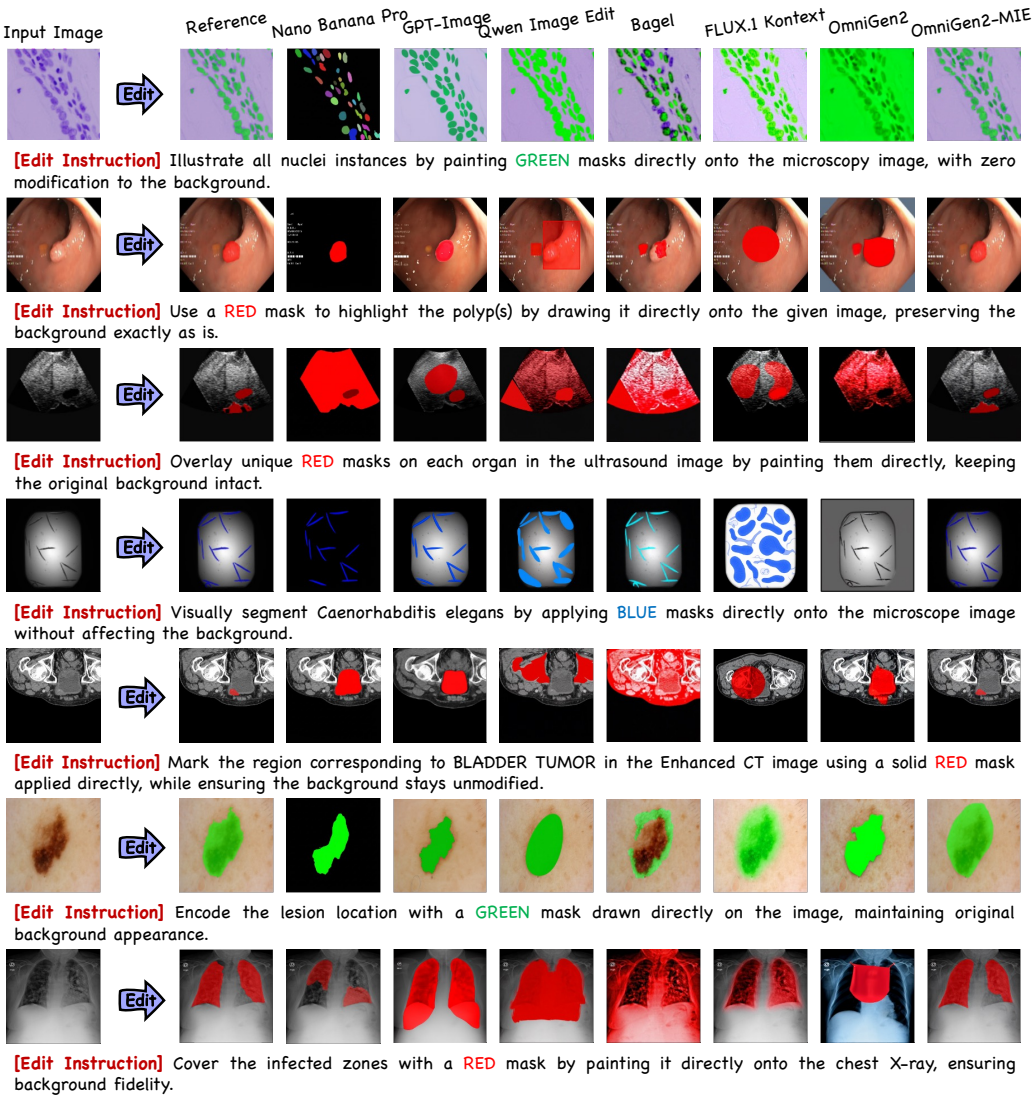
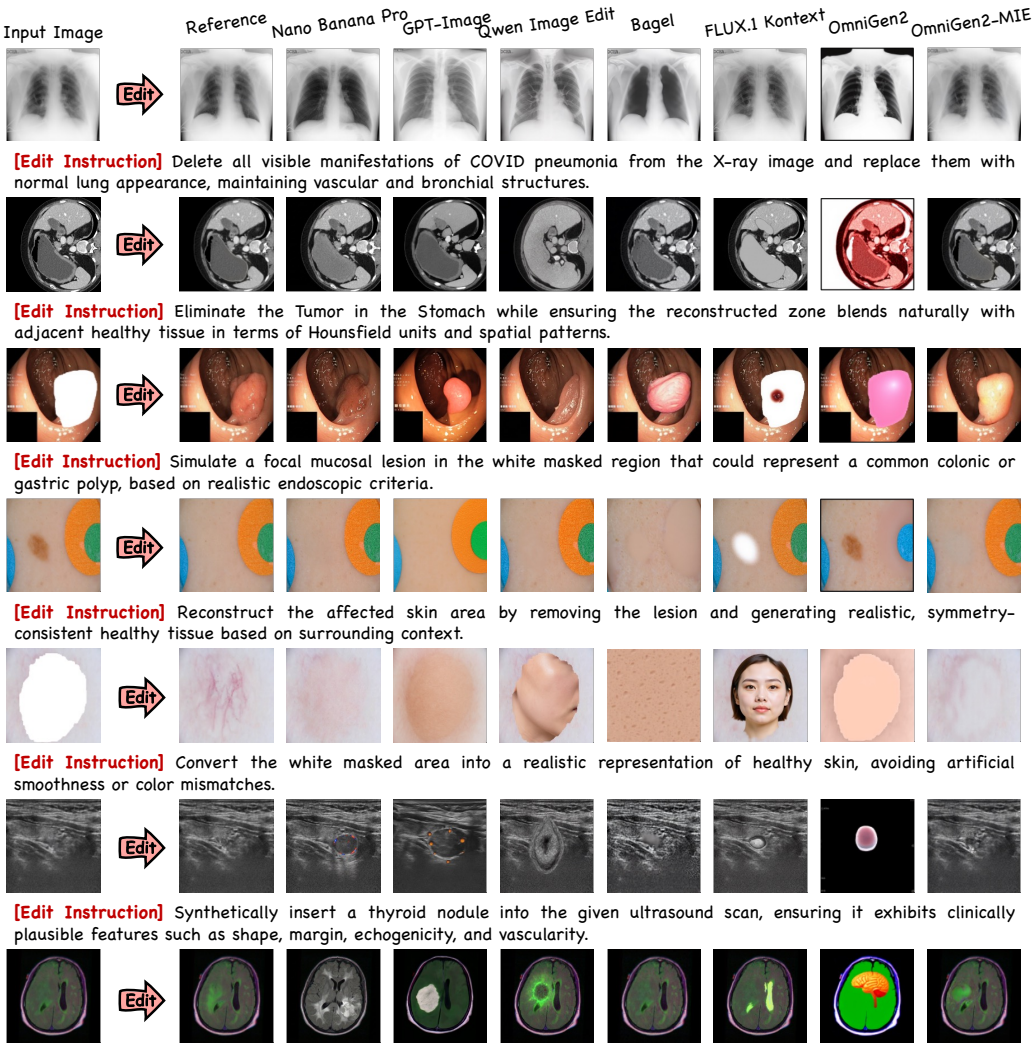
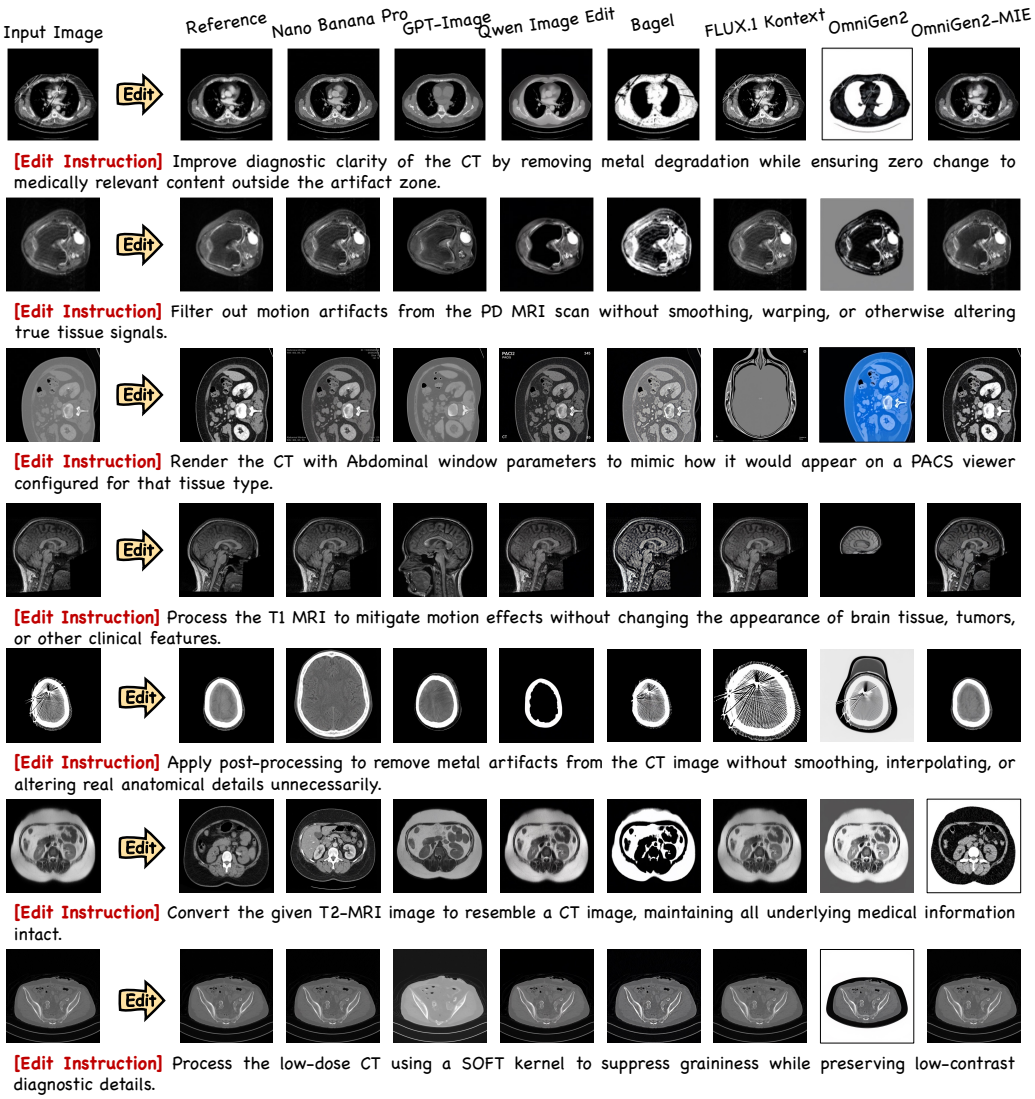


Figure 12: **Examples of Inpainting.** We train different inpainting models on each medical modality. H: the Healthy image; L: the Lesion-bearing image.

F MORE QUALITATIVE RESULTS




Input Image
Reference
Nano Banana Pro
GPT-Image
Qwen Image Edit
Bagel
FLUX.1 Kontext
OmniGen2
OmniGen2-MIE
Edit
[Edit Instruction] Delete all visible manifestations of COVID pneumonia from the X-ray image and replace them with normal lung appearance, maintaining vascular and bronchial structures.
[Edit Instruction] Eliminate the Tumor in the Stomach while ensuring the reconstructed zone blends naturally with adjacent healthy tissue in terms of Hounsfield units and spatial patterns.
[Edit Instruction] Simulate a focal mucosal lesion in the white masked region that could represent a common colonic or gastric polyp, based on realistic endoscopic criteria.
[Edit Instruction] Reconstruct the affected skin area by removing the lesion and generating realistic, symmetry-consistent healthy tissue based on surrounding context.
[Edit Instruction] Convert the white masked area into a realistic representation of healthy skin, avoiding artificial smoothness or color mismatches.
[Edit Instruction] Synthetically insert a thyroid nodule into the given ultrasound scan, ensuring it exhibits clinically plausible features such as shape, margin, echogenicity, and vascularity.
[Edit Instruction] Add a convincing brain tumor to the FLAIR MRI while maintaining correct contrast dynamics—such as suppressed CSF and bright pathological signal.


Input Image
Reference
Nano Banana Pro
GPT-Image
Qwen Image Edit
Bagel
FLUX.1 Kontext
OmniGen2
OmniGen2-MIE
Edit
[Edit Instruction] Improve diagnostic clarity of the CT by removing metal degradation while ensuring zero change to medically relevant content outside the artifact zone.
[Edit Instruction] Filter out motion artifacts from the PD MRI scan without smoothing, warping, or otherwise altering true tissue signals.
[Edit Instruction] Render the CT with Abdominal window parameters to mimic how it would appear on a PACS viewer configured for that tissue type.
[Edit Instruction] Process the T1 MRI to mitigate motion effects without changing the appearance of brain tissue, tumors, or other clinical features.
[Edit Instruction] Apply post-processing to remove metal artifacts from the CT image without smoothing, interpolating, or altering real anatomical details unnecessarily.
[Edit Instruction] Convert the given T2-MRI image to resemble a CT image, maintaining all underlying medical information intact.
[Edit Instruction] Process the low-dose CT using a SOFT kernel to suppress graininess while preserving low-contrast diagnostic details.

G FAILURE CASES

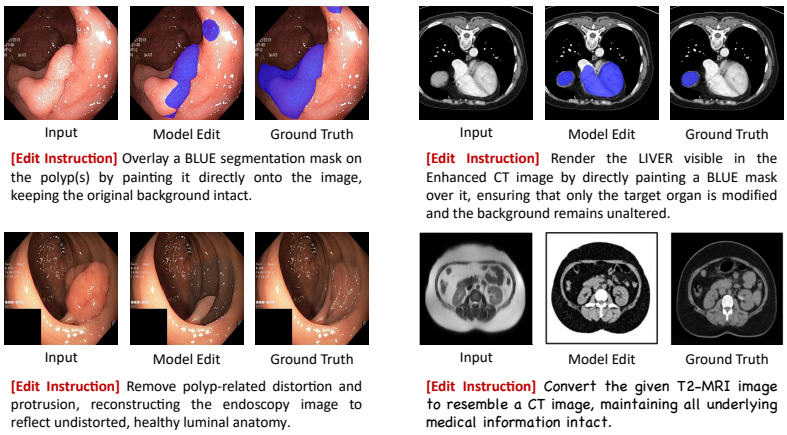


Figure 13: **Failure Cases.**

Fig 13 illustrates several representative failure cases of OmniGen2-MIE. The most frequent failure modes include: (1) semantic confusion between targeted anatomical features and morphologically similar background tissues; (2) intensity inconsistency, where the brightness of the edited region deviates from the surrounding context in a physically implausible manner; and (3) background inconsistency, especially after holistic transformations. These limitations underscore the need for more sophisticated multimodal architectures capable of preserving fine-grained details, as well as even more comprehensive training datasets to satisfy the requirements of rigorous clinical applications.

H LIMITATIONS AND BROADER IMPACTS

Although MieDB-100k provides a large-scale and diverse dataset for medical image editing, the primary limitation lies in the inherent difficulty of capturing ALL possible medical imaging modalities, and the relative scarcity of data for rare clinical cases. Continuous efforts to enrich these underrepresented categories will be vital for enhancing the dataset's diversity and effectiveness. Furthermore, while our work establishes a foundation for unified understanding and generation, it focuses exclusively on editing tasks. Integrating medical VQA and text-to-image datasets represents a natural progression of this research direction, resulting in a more comprehensive resource for the development of holistic medical models.

Our work has two potential societal risks. First, generative medical image editing could be misused to fabricate or alter clinical evidence, posing risks to patient safety and diagnostic integrity; we therefore explicitly state that this dataset is intended solely for research purposes and is **not suitable for direct clinical decision-making**. Second, biases present in the source datasets—such as underrepresentation of certain demographics, anatomical regions, or rare conditions—may propagate or be amplified by models trained on MieDB-100k, potentially leading to inequitable performance across patient populations. We encourage downstream users to perform thorough bias audits and to comply with applicable regulatory and ethical guidelines before deploying any models trained on this data in clinical or real-world settings.