# MedEBench: Revisiting Text-instructed Image Editing on Medical Domain

**Anonymous ACL submission**


## Abstract

Text-guided image editing has seen rapid progress in natural image domains, but its adaptation to medical imaging remains limited and lacks standardized evaluation. Clinically, such editing holds promise for simulating surgical outcomes, creating personalized teaching materials, and enhancing patient communication. To bridge this gap, we introduce **MedEBench**, a comprehensive benchmark for evaluating text-guided medical image editing. It consists of 1,182 clinically sourced image-prompt triplets spanning 70 tasks across 13 anatomical regions. MedEBench offers three key contributions: (1) a clinically relevant evaluation framework covering Editing Accuracy, Contextual Preservation, and Visual Quality, supported by detailed descriptions of expected change and ROI (Region of Interest) masks; (2) a systematic comparison of seven state-of-the-art models, revealing common failure patterns; and (3) a failure analysis protocol based on attention grounding, using IoU (Intersection over Union Ratio) between attention maps and ROIs to identify mislocalization. MedEBench provides a solid foundation for developing and evaluating reliable, clinically meaningful medical image editing systems.


## 1 Introduction

Recent advances in diffusion models and vision-language pretraining have significantly advanced text-guided image editing (Brooks et al., 2023a; Kawar et al., 2023a; Geng et al., 2023; Wasserman et al., 2025; Zhang et al., 2025a; Ge et al., 2025; DeepMind, 2024). These methods enable diverse applications such as object removal, inpainting, and style transfer, often producing compelling visual results (Yildirim et al., 2023; Wang et al., 2022; Yang et al., 2024). However, most prior work has focused on aesthetic or creative tasks, with limited exploration in domains requiring high semantic precision—particularly medicine. While comprehensive benchmarks exist for evaluating editing models on natural images (Lin et al., 2015; Huang et al., 2023; Xia et al., 2021), their adaptation to medical images remains challenging. Here, even minor alterations can carry significant clinical meaning, demanding high editing precision, semantic fidelity, and anatomical correctness. Text-guided medical image editing holds substantial clinical potential, as it could highlight lesions in CT scans (Guo et al., 2023), simulate surgical outcomes (Huang et al., 2025), or generate personalized teaching materials (Lee et al., 2024). Such applications promise direct benefits to diagnosis, treatment optimization (Ma, 2025), and training (Zhang et al., 2024a). Despite the versatility of current models, they often fail at clinically meaningful transformations that are intuitive to general physicians or even non-experts. For example, InstructPix2Pix (Brooks et al., 2023a) can successfully handles prompts like "Add a missing key on a piano keyboard" but fails on similar medical prompts such as "Add a missing tooth in this dental image" (see Fig. 1). Notably, this failure cannot be resolved by adjusting the text or image guidance scales (see Fig. 7). This underscores the difficulty of transferring such models to domains that require fine-grained anatomical under-

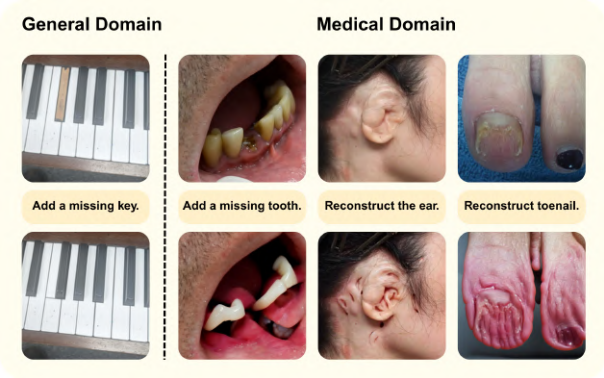



Figure 1: State-of-the-Art model performs well on **common images** (e.g., "add a missing key") but surprisingly struggles with **medical images** (e.g., "add a missing tooth").

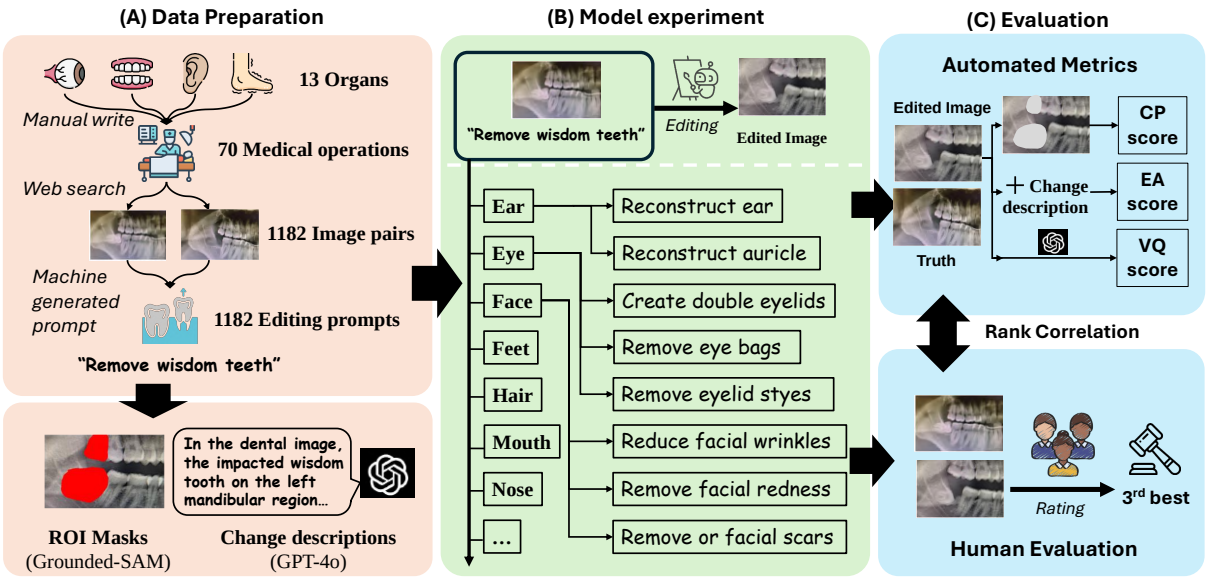



Figure 2: Overview of **MedEBench**, a text-guided benchmark for medical image editing. (A) Data preparation includes collecting image triplets, generating ROI masks, and describing intended changes. (B) Models generate edited images from prompts and previous images. (C) SSIM measures contextual preservation; editing accuracy and visual quality are assessed by GPT-4o, guided by the change description.

standing and specialized domain knowledge (Kazerouni et al., 2022; Ayana et al., 2024).

Medical image editing faces two fundamental challenges. First, real image pairs that reflect clinically grounded transformations (e.g., before and after treatment) are scarce, and synthetic data often lacks realism. Second, evaluation remains problematic. Metrics like Fréchet Inception Distance (FID) (Heusel et al., 2018) and CLIP Score (Hessel et al., 2022) are not aligned with clinical correctness or anatomical plausibility, failing to capture if the edits are grounded in medical knowledge.

To address these gaps, we propose **MedEBench**, a benchmark for text-guided medical image editing. MedEBench contains 1,182 real clinical image pairs covering pre- and postoperative states across 13 anatomical regions (e.g., teeth, eyes, esophagus). Each task is defined by a natural language prompt, region-of-interest (ROI) masks, and detailed change descriptions to enable localized, fine-grained evaluation. For clinically meaningful assessment, we introduce tailored evaluation metrics. *Contextual Preservation* (CP) is measured by masked SSIM (Wang et al., 2004) to ensure unaffected regions remain intact. *Editing Accuracy* (EA) and *Visual Quality* (VQ) are assessed via Multimodal Large Language Models (MLLMs) using clinically detailed change descriptions for accurate, interpretable evaluation. We benchmark seven state-of-the-art models across diverse learning paradigms and analyze failures through attention-grounding, revealing gaps in medical concept understanding and spatial localization. Our contributions are threefold:

- We introduce **MedEBench**, the first benchmark for text-guided medical image editing, featuring 1,182 real clinical image pairs from 13 anatomical regions and 70 editing tasks, each with prompts, ROI masks, and change descriptions.
- We propose clinically grounded evaluation metrics that capture structural preservation, edit accuracy, and visual quality, integrating both traditional approaches and MLLM-based reasoning.
- We benchmark seven models with text-instructed image editing ability on different learning paradigms and provide failure case analysis to diagnose limitations in medical concept editing by attention grounding.

## 2 MedEBench

Fig. 2 illustrates the main components of MedEBench. Section 2.1 describes our dataset, which consists of 1,182 samples across 13 categories. Each sample includes an input image, a reference image, an editing prompt, an ROI mask, and a change description. Section 2.2 introduces automated evaluation metrics, while Section 3.5 presents a human study validating the alignment between automated and human assessments. Tab. 1 compares MedEBench with existing benchmarks.

| Benchmark | Size | Domain | Synthetic | Truth |
|---|---|---|---|---|
| EditVal (Basu et al., 2023) | 648 pairs, 13 edit types | General | ✗ | ✗ |
| I2EBench (Ma et al., 2024) | 2000+ pairs, 16 edit types | General | ✗ | mix |
| EditBench (Wang et al., 2023b) | 240 pairs | General | mix | ✗ |
| PIE-Bench (Ju et al., 2023) | 700 pairs, 10 edit types | General | mix | ✗ |
| **MedEBench (Ours)** | 1182 pairs, 13 organs, 70 types | **Medical** | ✗ | ✓ |

Table 1: Comparison of text-guided image editing benchmark datasets.

As shown in Fig. 4, low prompt-image CLIP similarity and varied ROI sizes highlight the challenges of text-guided medical image editing.

## 2.1 Data Preparation

A major challenge in medical image editing is the scarcity of image pairs that reflect real clinical interventions. Unlike natural image datasets, synthesizing realistic transformations for medical images is difficult due to anatomical complexity and clinical constraints.

### 2.1.1 Image Pair Collection

To mitigate data scarcity, we curate a set of "before-and-after" medical image pairs simulating realistic clinical transformations across various anatomical regions. We define a list of target organs $\mathcal{O}$ (e.g., *Teeth*) and use ChatGPT (OpenAI, 2024) to generate corresponding medical procedures $\mathcal{A}_o$ (e.g., *Remove wisdom teeth*), as shown in Tab. 5 and Fig. 3. Guided by these organ-procedure pairs, two expert annotators collected image pairs $(I_{\text{prev}}, I_{\text{after}})$ via keyword-based web search. As no existing dataset met our quality and alignment needs, they screened large volumes of candidates for anatomical accuracy, procedural relevance, and visual consistency. All pairs underwent two-stage review, were standardized to 512-pixel width, aligned via affine transformations, and anonymized for privacy.

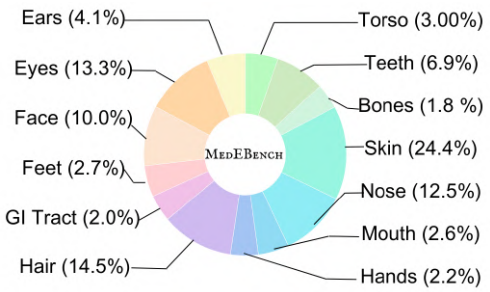



Figure 3: Organ distribution

### 2.1.2 Editing Prompt Generation

For each verified image pair $(I_{\text{prev}}, I_{\text{after}})$, we manually create a natural language prompt $p_{\text{orig}}$ describing the visual transformation, concisely reflecting the underlying medical operation. To enhance linguistic diversity and mitigate overfitting to fixed

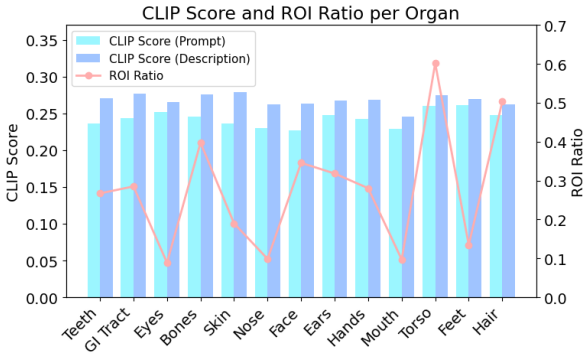



Figure 4: Properties of the MedEBench dataset: CLIP score between the instruction and the preceding image, and the ROI mask ratio relative to the full image.

phrasings, we use ChatGPT (OpenAI, 2024) to generate multiple paraphrased variants. One variant is randomly selected as $p_{\text{reph}}$ for use in evaluation.

### 2.1.3 Region-of-Interest Mask Annotation

For each image $I_{\text{prev}}$, we generate a region-of-interest (ROI) mask $M$ to localize the area targeted by the editing prompt. Candidate masks are produced by prompting Grounded-SAM (Ren et al., 2024) with the instruction, yielding three proposals. A human annotator selects the most accurate mask or manually refines it to ensure anatomical precision. ROI masks serve two key purposes: (1) **Contextual Preservation**, where SSIM (Wang et al., 2004) is computed outside the masked region to verify that unedited areas remain unchanged; and (2) **Attention Analysis**, which evaluates whether model attention aligns with the relevant anatomy.

### 2.1.4 Description of Change Generation

Clinically meaningful evaluation requires more than visual fidelity or prompt alignment—it demands a precise understanding of anatomical changes. To this end, we generate structured *descriptions of change* for each editing task, specifying the target anatomy and expected post-edit outcome. These task-specific descriptions replace generic prompts and guide Medical Large Language Model (MLLM)-based evaluations, enhancing interpretability and reliability. We employ GPT-4o (OpenAI, 2024) for its strong visual reasoning. For each sample, GPT-4o takes the input image

pair and editing instruction, and generates a brief summary of the intended anatomical change. These descriptions support both automated and qualitative evaluation. See Appendix H.3 for the full prompt.

### 2.2 Automated Evaluation

We develop an automated protocol to evaluate model performance across *Contextual Preservation*, *Editing Accuracy*, and *Visual Quality*, combining traditional image metrics with GPT-4o-based judgment. A more detailed explanation and demonstration are provided in the Appendix.

**Contextual Preservation** To evaluate whether the model preserves image regions unrelated to the intended edit, we compute SSIM(Wang et al., 2004) between the previous image $I_{\text{prev}}$ and the edited image $I_{\text{edit}}$, excluding the region-of-interest (ROI) mask $\mathcal{R}$. The contextual SSIM is defined as: $\text{SSIM}_{\text{context}} = \text{SSIM}\left(I_{\text{prev}}|_{\overline{\mathcal{R}}}, I_{\text{edit}}|_{\overline{\mathcal{R}}}\right)$. This metric captures how well the model maintains anatomical consistency outside the edited region.

**Editing Accuracy and Visual Quality** We use GPT-4o (OpenAI, 2024), a multimodal large language model with visual reasoning capabilities, to evaluate *Editing Accuracy* and *Visual Quality*. For each sample, GPT-4o is provided with the description of change, previous image, edited image, and ground truth image, and follows a structured two-step protocol:

**Step 1: Visual Difference Description.** GPT-4o first compares the previous and edited images to describe all visible changes, identifying what has been added, removed, or modified, along with the anatomical regions affected.

**Step 2: Scoring.** Guided by the reference *description of change* generated for the task, GPT-4o evaluates the following aspects: **Editing Accuracy (0–10)** measures how well the actual changes in the edited image match the expected transformation described in the reference, reflecting completeness and correctness with deductions for irrelevant or missing edits; **Visual Quality (0–10)** assesses the realism, clarity, and overall visual fidelity of the edited image. Each score is accompanied by a concise rationale to enhance transparency and evaluation reliability.

## 3 Experiment and Results

### 3.1 Baseline Models

We evaluate seven state-of-the-art models with text-guided image editing capabilities. TIE models: 1) **InstructPix2Pix** (Brooks et al., 2023a): An early diffusion-based model fine-tuned on synthetic instruction-image pairs for prompt-based editing with strong spatial alignment. 2) **Imagic** (Kawar et al., 2023a): Optimizes latent codes to enable realistic edits of real images, preserving identity and structure without requiring paired data. 3) **InstructDiffusion** (Geng et al., 2023): A generalist model for instruction-following across diverse vision tasks, supporting flexible zero-shot editing. 4) **Paint-by-Inpaint** (Wasserman et al., 2025): Proposes an object addition paradigm via region removal and inpainting-based completion, enabling mask-free object insertion. 5) **ICEdit** (Zhang et al., 2025a): Utilizes diffusion transformers (DiT) with in-context learning and adapter tuning for few-shot instructional editing. Universal MLLMs: 6) **SEED-X** (Ge et al., 2025): A unified multimodal model supporting both image understanding and generation for general-purpose editing. 7) **Gemini 2.0 Flash** (DeepMind, 2024): A commercial-grade multimodal system integrating fast image generation, conversational interaction, and robust editing capabilities. To ensure fair comparison, we perform hyperparameter sweeps for each model around default configurations. Detailed settings and prompts are provided in the Appendix.

### 3.2 Baseline Metrics

We compare our proposed metrics (Section 2.2) with commonly used automated metrics for image editing quality assessment, focusing on their correlation with human evaluation results (Section 3.5). For CLIP-based metrics: 1) **ISim** measures the similarity between the edited image and the ground truth in the CLIP embedding space; 2) **TAlign** evaluates the alignment between text instructions and edited images; 3) **DAlign** captures the directional consistency of edits with respect to text guidance. In the Pixel Similarity category, traditional image quality metrics compare edited images against ground truth: 4) **PSNR** (Korhonen and You, 2012) measures pixel-wise reconstruction accuracy; 5) **LPIPS** (Zhang et al., 2018) quantifies perceptual similarity using deep visual features; and 6) **SSIM** (Wang et al., 2004) assesses structural similarity. Additionally, the reward-based metric 7) **ImageReward** (Xu et al., 2023) provides a learned perceptual score designed to correlate with human preferences. Finally, 8) **FID** (Heusel et al., 2018) measures the distance between the distributions of real and generated images in a feature space. Since FID is computed at the distribution level, it

| Model | Teeth | | | Eyes | | | Spine | | | Skin | | | Nose | | | Face | | | GI Tract | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ |
| imagic | 0.49 | 0.64 | 0.57 | 0.50 | 0.49 | 0.62 | 0.18 | 0.66 | 0.52 | 0.19 | 0.53 | 0.37 | 0.40 | 0.40 | 0.48 | 0.27 | 0.55 | 0.42 | 0.13 | 0.63 | 0.58 |
| instruct-pix2pix | 0.32 | **0.85** | 0.42 | 0.59 | **0.90** | **0.75** | 0.21 | 0.79 | 0.36 | 0.51 | **0.85** | 0.65 | 0.66 | **0.86** | 0.72 | 0.70 | 0.85 | **0.79** | 0.19 | **0.87** | 0.67 |
| instruct-diffusion | 0.47 | 0.67 | 0.65 | **0.64** | 0.66 | 0.72 | 0.06 | 0.72 | 0.52 | 0.56 | 0.69 | 0.59 | 0.59 | 0.57 | 0.69 | 0.54 | 0.68 | 0.58 | 0.18 | 0.62 | 0.45 |
| paint-by-inpaint | 0.19 | 0.67 | 0.25 | 0.40 | 0.57 | 0.43 | 0.14 | 0.67 | 0.27 | 0.25 | 0.47 | 0.30 | 0.54 | 0.61 | 0.54 | 0.36 | 0.54 | 0.36 | 0.10 | 0.57 | 0.33 |
| icedit | 0.46 | 0.79 | 0.57 | 0.60 | 0.72 | 0.69 | 0.16 | 0.67 | 0.50 | 0.64 | 0.78 | 0.72 | 0.72 | 0.77 | **0.75** | 0.71 | 0.78 | 0.76 | 0.37 | 0.78 | 0.70 |
| seedx | 0.32 | 0.70 | 0.47 | 0.33 | 0.79 | 0.55 | 0.37 | 0.65 | 0.48 | 0.26 | 0.79 | 0.54 | 0.24 | 0.77 | 0.57 | 0.33 | 0.76 | 0.58 | 0.22 | 0.61 | 0.49 |
| gemini_2_flash | **0.75** | 0.84 | **0.81** | 0.55 | 0.78 | 0.63 | **0.38** | **0.81** | **0.76** | **0.77** | 0.84 | **0.77** | **0.77** | 0.76 | 0.72 | **0.72** | **0.86** | 0.70 | **0.66** | 0.80 | **0.76** |

| Ears | | | Hands | | | Mouth | | | Torso | | | Feet | | | Hair | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | EA | CP | VQ | FID |
| 0.40 | 0.58 | 0.50 | 0.41 | 0.51 | 0.45 | 0.40 | 0.43 | 0.39 | 0.33 | 0.67 | 0.51 | 0.26 | 0.55 | 0.38 | 0.64 | 0.71 | 0.53 | 0.38 | 0.55 | 0.48 | 88.54 |
| 0.36 | **0.93** | 0.66 | 0.36 | 0.75 | 0.39 | 0.64 | 0.77 | 0.64 | 0.18 | 0.73 | 0.35 | 0.17 | **0.90** | 0.48 | 0.42 | 0.88 | 0.44 | 0.50 | **0.86** | 0.62 | 46.50 |
| 0.45 | 0.75 | 0.63 | 0.49 | 0.66 | 0.59 | 0.61 | 0.51 | 0.63 | 0.31 | 0.67 | 0.54 | 0.23 | 0.75 | 0.74 | 0.68 | 0.70 | 0.65 | 0.54 | 0.67 | 0.63 | 65.56 |
| 0.25 | 0.66 | 0.33 | 0.40 | 0.67 | 0.40 | 0.41 | 0.59 | 0.43 | 0.35 | 0.66 | 0.30 | 0.16 | 0.58 | 0.20 | 0.60 | 0.71 | 0.51 | 0.37 | 0.59 | 0.38 | 109.61 |
| 0.70 | 0.83 | 0.76 | 0.50 | 0.70 | 0.68 | 0.71 | 0.71 | 0.74 | 0.62 | **0.90** | **0.74** | 0.74 | 0.84 | 0.76 | **0.71** | **0.90** | **0.79** | 0.60 | 0.79 | **0.72** | 46.76 |
| 0.41 | 0.81 | 0.51 | 0.43 | 0.74 | 0.54 | 0.27 | **0.80** | 0.67 | 0.35 | 0.72 | 0.50 | 0.07 | 0.88 | 0.56 | 0.40 | 0.86 | 0.52 | 0.31 | 0.78 | 0.54 | 51.04 |
| **0.76** | 0.72 | **0.82** | **0.82** | **0.81** | **0.79** | **0.73** | 0.72 | **0.79** | **0.62** | 0.87 | **0.74** | **0.82** | 0.86 | **0.82** | 0.50 | 0.87 | 0.63 | **0.68** | 0.82 | **0.72** | **43.74** |

Table 2: Editing performance across organs. **Bold** = best, underline = second best. EA = GPT-4o Editing Accuracy$_{\text{Desc}}$, CP = Masked SSIM, VQ = GPT-4o Visual Quality$_{\text{Desc}}$. EA and VQ scaled from 0–10 to 0–1.

is not included in our correlation comparison but is reported as a reference in Section 3.3.

### 3.3 Main Results

We summarize model performance on MedEBench in Tab. 2 and visualize key trends in Fig. 5. Representative editing examples are shown in Tab. 6. Gemini 2 Flash achieves the best overall performance, leading in editing accuracy (EA = 0.68), visual quality (VQ = 0.72, tied with ICEdit), and ranking second in context preservation (CP = 0.82). It also delivers the most realistic outputs, as reflected by its superior Fréchet Inception Distance (FID = 43.74) (Heusel et al., 2018). Among open-source models, ICEdit shows the most balanced performance (EA = 0.60, VQ = 0.72, CP = 0.79). InstructPix2Pix, while excelling in context preservation (CP = 0.86), suffers from lower editing accuracy (EA = 0.50), likely due to its reluctance to apply medically significant edits. A substantial gap persists between Gemini and open-source methods with over 10% in editing accuracy, highlighting limitations of current methods for high-precision medical editing.

**At the organ level**, editing difficulty varies significantly. Regions like the *spine* and *bones* (e.g., CT scans) remain challenging, with the best EA only reaching 0.38. The largest performance gaps between Gemini and open-source models are seen in *teeth* (34%), *hands* (64%), and *gastrointestinal tract* (43%), reflecting the complexity of anatomical structures and repetitive patterns. Conversely, superficial structures like *hair* and *nose* are easier to edit, where ICEdit achieves strong performance (hair: EA = 0.71, VQ = 0.79).

**Model-specific strengths** are observed. Gemini performs well on precision-demanding edits and internal organ modifications, though its context preservation slightly trails InstructPix2Pix. ICEdit, powered by DiT-based in-context editing, handles complex anatomy effectively but struggles with concept removal (e.g., *teeth* edits). SEED-X shows surprising strength in *spine* edits (EA = 0.37), outperforming other open-source models, possibly due to better alignment with CT image modalities.

Three key observations arise: **(1) Internal organs** are significantly harder to edit than superficial structures, with EA scores averaging 2.3× lower; **(2) Gemini's advantage** is most pronounced in anatomically complex regions with repetitive patterns (e.g., hands, gastrointestinal tract); and **(3) A trade-off exists** between visual quality and context preservation: models with higher VQ, such as Gemini, tend to exhibit slightly lower CP, highlighting the need for localized, precise edits rather than full-image redraws in medical applications.

### 3.4 Learning Paradigms Comparison

To explore different learning paradigms for medical image editing, we select six representative tasks and sample 30 images per task. We compare two representative approaches: **Fine-tuning.** We adopt InstructPix2Pix as a representative fine-tuning method. Its U-Net backbone is fine-tuned using triplets of input images, editing instructions, and ground-truth edited outputs. Training is per-

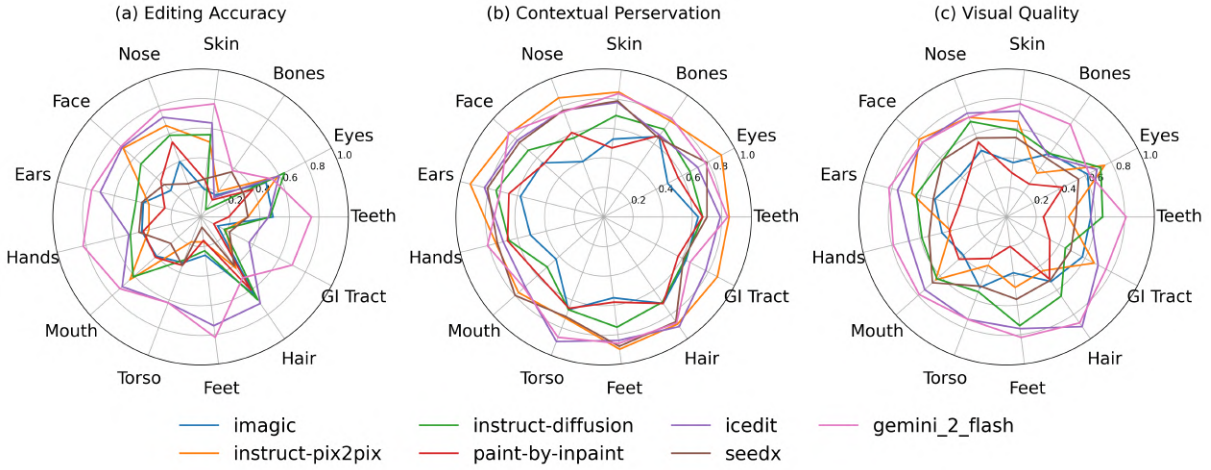



Figure 5: Per-organ performance comparison of seven image editing models across three metrics.

formed for 50 steps per sample. **In-context learning.** We evaluate Gemini's in-context capability via few-shot prompting. During inference, prompts are constructed using several demonstration triplets (previous image, editing instruction, and ground-truth), followed by a test image and instruction.

As shown in Tab. 7, fine-tuning proves effective for InstructPix2Pix, particularly in reconstruction tasks. Editing accuracy improves consistently with more fine-tuning samples. However, tasks like *Remove Wisdom Teeth* and *Remove Moles* show diminishing returns, with accuracy gains plateauing and often compromising contextual preservation.

In contrast, Gemini's in-context learning shows limited effectiveness. Increasing the number of demonstrations does not improve accuracy and often degrades performance. The model struggles to distinguish between test inputs and in-context examples, leading to confusion and reduced contextual consistency. These findings highlight the difficulty of applying in-context learning to fine-grained, pixel-level medical editing.

### 3.5 Human Evaluation

To validate our automated metrics, we conduct a human evaluation study based on relative ranking. Two expert annotators independently assess a subset of edited images across three dimensions: *Editing Accuracy* (EA), *Contextual Preservation* (CP), and *Visual Quality* (VQ), consistent with the automated evaluation framework. The evaluation panel and annotator instructions appear in Fig. 10. For each organ $o \in \mathcal{O}$, we randomly sample 20 editing tasks, yielding a total of 260 edited samples. Outputs from multiple models are collected using identical inputs and prompts. Annotators

| | Metric | EA | CP | VQ |
|---|---|---|---|---|
| CLIP | ISim | 0.35 | 0.55 | 0.56 |
| | $\text{TAlign}_\text{P}$ | 0.16 | 0.26 | 0.28 |
| | $\text{TAlign}_\text{D}$ | 0.22 | 0.36 | 0.33 |
| | $\text{DAlign}_\text{P}$ | 0.15 | 0.25 | 0.24 |
| | $\text{DAlign}_\text{D}$ | 0.20 | 0.35 | 0.32 |
| Pixel/Reward | PSNR | 0.33 | 0.65 | 0.54 |
| | LPIPS | 0.38 | 0.67 | 0.60 |
| | SSIM | 0.20 | 0.66 | 0.43 |
| | **Masked SSIM** | 0.21 | **0.82** | 0.45 |
| | $\text{ImageReward}_\text{P}$ | 0.20 | 0.28 | 0.41 |
| | $\text{ImageReward}_\text{D}$ | 0.16 | 0.27 | 0.36 |
| GPT-4o | $\text{Accuracy}_\text{P}$ | 0.64 | 0.21 | 0.42 |
| | $\textbf{Accuracy}_\textbf{D}$ | **0.79** | 0.18 | 0.46 |
| | $\text{Context}_\text{P}$ | 0.41 | 0.56 | 0.69 |
| | $\text{Context}_\text{D}$ | 0.43 | 0.57 | 0.76 |
| | $\text{Quality}_\text{P}$ | 0.44 | 0.44 | 0.81 |
| | $\textbf{Quality}_\textbf{D}$ | 0.46 | 0.46 | **0.82** |

Table 3: Spearman Rank Correlation ($\rho$) Between Human Ratings (column) and Automated Metrics (row).

rank the model outputs for each sample along all three dimensions. The inter-annotator Spearman rank correlation coefficient reaches 0.91, indicating excellent agreement and high annotation reliability.

To assess the consistency between human and automated evaluations, we compute the Spearman correlation coefficient $\rho$ (Spearman, 1904) between human-assigned ranks $R_h(i)$ and automated ranks $R_a(i)$ for each sample $i$ across $m$ models:

$$\rho = 1 - \frac{6 \sum_i (R_h(i) - R_a(i))^2}{m(m^2 - 1)} \quad (1)$$

We evaluate both baseline metrics (Section 3.2) and our proposed ones (Section 2.2). To further investigate the impact of prompt structure, we introduce alternative versions of each text-based metric by replacing the original editing prompts with structured *descriptions of change* (Section 2.1.4). As shown in

Tab. 3, our proposed metrics (Masked SSIM for CP, GPT-4o based EA, and VQ with detailed change descriptions) achieve the highest alignment with human assessments across all three evaluation dimensions. Notably, incorporating expected change guidance into GPT-4o scoring improves correlation with expert judgments by 23%.

## 4 Failure Analysis via Attention

We take the InstructPix2Pix model as a case study to analyze failure cases. Notably, as discussed in Section 3.3, InstructPix2Pix demonstrates high context preservation but relatively low editing accuracy. As shown in (Liu et al., 2024a), cross-attention helps localize prompt tokens to image regions and carries semantic and categorical information. To investigate this further, we examine whether the editing process attends to the correct anatomical regions by analyzing the model's cross-attention maps. For each editing prompt, we extract the cross-attention maps corresponding to the last token of the key visual concept $t_c$ across all diffusion steps. These are averaged to obtain a single attention vector: $\bar{\mathbf{a}}_{t_c} = \frac{1}{S}\sum_{s=1}^{S} \mathbf{a}_{t_c}^{(s)}$. This vector is reshaped into a 2D map $\bar{A}_{t_c}$, normalized to the range $[0, 1]$, and binarized using a scaled (We choose $\alpha = 1.3$.) Otsu threshold (Otsu, 1979):

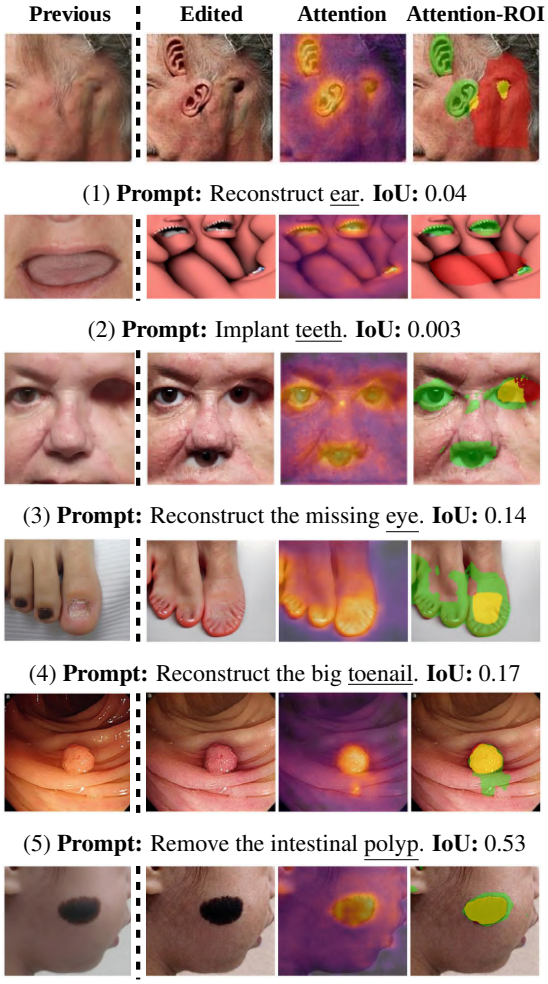
$$A = \left( \frac{\bar{A}_{t_c} - \min}{\max - \min} > \alpha \cdot \tau_{\text{Otsu}} \right) \tag{2}$$

We then compare the thresholded attention map $A$ with the annotated Region-of-Interest (ROI) mask $M$ using the Intersection-over-Union: $\text{IoU} = |A \cap M| / |A \cup M|$. A higher IoU score indicates stronger spatial alignment between the model's attention and the intended anatomical target, reflecting better grounding of visual concepts during editing. The average IoU scores for each organ are listed in Tab. 4.

| Organ | IoU | Organ | IoU | Organ | IoU |
|---|---|---|---|---|---|
| Feet | 0.189 | Teeth | 0.268 | GI Tract | 0.4084 |
| Skin | 0.191 | Face | 0.289 | Hands | 0.409 |
| Nose | 0.193 | Bones | 0.389 | Hair | 0.497 |
| Eyes | 0.195 | Ears | 0.407 | Torso | 0.592 |
| Mouth | 0.224 | - | - | - | - |

Table 4: Average IoU between attention maps and ROI masks across anatomical regions in InstructPix2Pix.

**Analysis** We analyze failure cases characterized by **low Editing Accuracy (EA)** and **high Context Preservation (CP)**, indicating that edits were insufficient or absent, though the overall image remained intact. In this regime, IoU between attention maps and ground-truth regions serves as a diagnostic indicator to distinguish failure types:




(1) **Prompt:** Reconstruct ear. **IoU:** 0.04

(2) **Prompt:** Implant teeth. **IoU:** 0.003

(3) **Prompt:** Reconstruct the missing eye. **IoU:** 0.14

(4) **Prompt:** Reconstruct the big toenail. **IoU:** 0.17

(5) **Prompt:** Remove the intestinal polyp. **IoU:** 0.53

(6) **Prompt:** Remove birthmark. **IoU:** 0.70

Figure 6: InstructPix2Pix cross-attention on key visual concept versus ground-truth ROI. In the last column: red = ROI mask, green = attention, yellow = overlap.

- **High IoU**: The model correctly localizes the target region but fails to apply the intended edit. This reflects partial spatial understanding with conservative or limited editing capability, often seen in concept removal tasks (e.g., samples 5–6 in Fig. 6).
- **Low IoU**: The model fails both to localize and to edit, attending to irrelevant regions. Such misalignment typically occurs in addition or reconstruction tasks, especially in anatomically complex areas (e.g., samples 1–4 in Fig. 6).

These patterns reveal fundamental limitations in current models' spatial reasoning and medical concept grounding, underscoring the challenges of reliable medical image editing.

## 5 Key Insights and Takeaways

Our comprehensive evaluation across thirteen anatomical regions and multiple editing models yields four key findings:

- **Large multimodal models outperform open-source alternatives on complex medical edits.** Gemini 2 Flash consistently leads in EA, VQ, and FID, with the largest gaps in internal organ tasks (e.g., gastrointestinal tract, spine, teeth) where fine-grained structures and spatial reasoning are critical. Open-source models like ICEdit remain competitive but lag in high-precision scenarios.
- **Region-specific challenges persist.** Models struggle with repetitive or occluded anatomy (e.g., hands, spine, teeth), while tasks involving superficial structures (e.g., skin blemishes, nose shape) are more reliably handled.
- **Fine-tuning aids domain adaptation; in-context learning shows limited generalization.** Fine-tuning models like InstructPix2Pix improves medical editing performance despite data scarcity. However, large multimodal models struggle to transfer medical concepts through prompting-level in-context learning, highlighting generalization limits in clinical tasks.
- **Attention maps provide diagnostic insights.** IoU between attention heatmaps and ground-truth regions reveals failure patterns: high IoU with low EA indicates correct localization but failed execution (common in concept removal), while low IoU reflects poor spatial focus in addition or reconstruction tasks.

## 6 Related Works

**Text-guided Image Editing** has advanced rapidly with diffusion-based models (Ho et al., 2020; Dhariwal and Nichol, 2021; Rombach et al., 2022; Saharia et al., 2022; Nichol et al., 2022; Ho et al., 2021; Zhang et al., 2023a; Meng et al., 2022; Ramesh et al., 2022; Chen et al., 2025), enabling natural language-conditioned image synthesis (Brooks et al., 2023b; Kawar et al., 2023b; Hertz et al., 2022; Mokady et al., 2022; Choi et al., 2023; Ravi et al., 2023; Kim et al., 2022; Nguyen et al., 2024; Zhang et al., 2025b; Wang et al., 2023a). Recent works (Kawar et al., 2023a; Brooks et al., 2023a; Geng et al., 2023; Wasserman et al., 2025; Zhang et al., 2025a; Ge et al., 2025; DeepMind, 2024) further improve control, generalization, and real-world applicability. Medical image editing is an emerging application, supporting tasks such as disease progression simulation (Taylor et al., 2019; Puglisi et al., 2024; Cao et al., 2024; Alaya et al., 2024), segmentation (Feng, 2024; Ma, 2025; Dong et al., 2024; Wu et al., 2023), and synthetic data generation (Cho et al., 2024; Zhang et al., 2023b, 2024c; Kidder et al., 2024; Li et al., 2024b).

**Benchmarking Text-Guided Editing** is key to evaluating model performance, but standard metrics like FID (Heusel et al., 2018), CLIP Score (Hessel et al., 2022), PSNR (Korhonen and You, 2012), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) fail to capture fine-grained edit quality and semantic intent. Recent benchmarks (Ma et al., 2024; Basu et al., 2023; Wang et al., 2023b; Ju et al., 2023) improve on this with human-aligned dimensions and diverse edit types. Multimodal Large Language Models (MLLMs), such as GPT-4o (OpenAI, 2024) and Gemini 2.5 Pro (Google DeepMind, 2025), have emerged as strong evaluators for visual and semantic alignment, enabling interpretable and human-aligned assessment frameworks (Li et al., 2024a; Fan et al., 2024; Jin et al., 2024; Zhang et al., 2024b; Liu et al., 2024b; Chen et al., 2023). However, existing work largely targets general-domain imagery, overlooking domain-specific needs like medical editing. To bridge this gap, we propose **MedEBench**, the first benchmark for text-guided medical image editing with a clinically grounded evaluation framework.

## 7 Conclusion

We present **MedEBench**, the first comprehensive benchmark for text-guided medical image editing, featuring 1,182 real-world examples across 13 anatomical regions and 70 clinically meaningful tasks. Each sample includes region-of-interest masks and expert-written change descriptions, enabling evaluation along *Editing Accuracy*, *Context Preservation*, and *Visual Quality*. Our evaluation of seven state-of-the-art models reveals notable limitations, especially for internal organs and anatomically complex regions. While Gemini 2 Flash leads overall, ICEdit emerges as the strongest open-source model. Cross-attention analyses expose persistent issues in spatial reasoning and anatomical grounding, with models often mislocalizing edits despite accurate prompts. These findings highlight the gap between general-purpose editors and the demands of medical image manipulation. MedEBench provides a foundation for developing and rigorously evaluating models tailored to safety-critical medical applications.

## Limitations

While MedEBench provides a comprehensive benchmark for text-guided medical image editing, several limitations remain. First, the benchmark focuses exclusively on editing tasks that correspond to real-world surgical or clinical operations. As a result, it does not cover more speculative or exploratory editing tasks that may be of interest for medical education, simulation, or synthetic data augmentation. This focus ensures clinical relevance but limits the diversity of task types included.

Second, current text-to-mask models, such as Grounded-SAM, face significant challenges when applied to medical images. Specifically, these models struggle to accurately generate region-of-interest (ROI) masks for anatomical structures without clear boundaries or with ambiguous visual features. In such cases, automatic mask generation often fails, necessitating human intervention to ensure anatomical precision.

## Ethics Statement

MedEBench is constructed using publicly available and de-identified medical images. To ensure transparency and reproducibility, the original URLs of all raw images are provided as part of the dataset release. No personally identifiable information (PII) or sensitive patient data is included. The benchmark strictly serves research purposes and does not support diagnostic, therapeutic, or clinical decision-making applications. Although editing tasks are designed to correspond to real-world surgical procedures, the benchmark does not promote automated clinical image generation without human expert oversight, and any misuse for unauthorized medical applications is strongly discouraged. In cases where text-to-mask models like Grounded-SAM fail to generate reliable masks, human annotators with medical expertise are employed to ensure anatomical accuracy under proper consent and contractual agreements. Recognizing the potential for dataset-induced biases, we strive for diversity across anatomical regions and clinical tasks, while acknowledging that further work is needed to address bias mitigation in synthetic image generation. The benchmark, models, and code are released solely for academic and non-commercial use under an appropriate open-source license, with a strict reminder that MedEBench is not a substitute for clinical judgment or professional medical practice.

**Potential Risks.** Despite the intended academic use of MedEBench, there are inherent risks associated with the misuse of medical image editing models. One primary concern is the generation of misleading or fabricated clinical images that could potentially be used in malicious contexts, such as misinformation, fraudulent clinical documentation, or unauthorized patient record manipulation. Additionally, the use of generative models trained on limited or biased datasets may inadvertently reinforce existing healthcare disparities if applied to populations not well represented in the benchmark. Another risk lies in the potential over-reliance on automated editing tools without sufficient medical expertise, which could lead to clinically inaccurate or unsafe modifications. To mitigate these risks, we emphasize responsible usage under expert supervision, explicit academic licensing terms, and continuous community oversight.

## References


Malek Ben Alaya, Daniel M. Lang, Benedikt Wiestler, Julia A. Schnabel, and Cosmin I. Bercea. 2024. Mededit: Counterfactual diffusion-based image editing on brain mri. *Preprint*, arXiv:2407.15270.

Getachew Ayana, Kibret Dese, Asamene M Abagaro, Kye-Chul Jeong, Seung-Dae Yoon, and Se-Woon Choe. 2024. Multistage transfer learning for medical images. *Artificial Intelligence Review*, pages 1–47.

Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. 2023. Editval: Benchmarking diffusion based text-guided image editing methods. *Preprint*, arXiv:2310.02426.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023a. Instructpix2pix: Learning to follow image editing instructions. *Preprint*, arXiv:2211.09800.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023b. Instructpix2pix: Learning to follow image editing instructions. *Preprint*, arXiv:2211.09800.

Xu Cao, Kaizhao Liang, Kuei-Da Liao, Tianren Gao, Wenqian Ye, Jintai Chen, Zhiguang Ding, Jianguo Cao, James M. Rehg, and Jimeng Sun. 2024. Medical video generation for disease progression simulation. *Preprint*, arXiv:2411.11943.

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

Yixiong Chen, Li Liu, and Chris Ding. 2023. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *Preprint*, arXiv:2305.10843.

Joseph Cho, Mrudang Mathur, Cyril Zakka, Dhamanpreet Kaur, Matthew Leipzig, Alex Dalal, Aravind Krishnan, Eubee Koo, Karen Wai, Cindy S. Zhao, et al. 2024. Medisyn: A generalist text-guided latent diffusion model for diverse medical image synthesis. *arXiv preprint arXiv:2405.09806*.

Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. 2023. Custom-edit: Text-guided image editing with customized diffusion models. *Preprint*, arXiv:2305.15779.

Google DeepMind. 2024. Gemini 2.0 and gemini flash experimental. Accessed: 2025-03-05.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis. *Preprint*, arXiv:2105.05233.

Zhiwei Dong, Genji Yuan, Zhen Hua, and Jinjiang Li. 2024. Diffusion model-based text-guided enhancement network for medical image segmentation. *Expert Systems with Applications*, 239:123549.

Zhiyuan Fan, Weinong Wang, Xing W, and Debing Zhang. 2024. SedarEval: Automated evaluation using self-adaptive rubrics. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16916–16930, Miami, Florida, USA. Association for Computational Linguistics.

Chun-Mei Feng. 2024. Enhancing label-efficient medical image segmentation with text-guided diffusion models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer Nature Switzerland.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2025. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *Preprint*, arXiv:2404.14396.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. 2023. Instructdiffusion: A generalist modeling interface for vision tasks. *Preprint*, arXiv:2309.03895.

Google DeepMind. 2025. Gemini 2.5 pro: Google's most advanced ai model. Accessed: 2025-05-13.

Yucheng Guo, Yifan Wang, Yifan Zhang, Yizhou Wang, and Xiaoxiao Li. 2023. Generatect: Text-conditional generation of 3d chest ct volumes. *arXiv preprint arXiv:2305.16037*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *Preprint*, arXiv:2208.01626.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Preprint*, arXiv:1706.08500.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. 2021. Cascaded diffusion models for high fidelity image generation. *Preprint*, arXiv:2106.15282.

Gexin Huang, Ruinan Jin, Yucheng Tang, Can Zhao, Tatsuya Harada, Xiaoxiao Li, and Gu Lin. 2025. Interactive tumor progression modeling via sketch-based image editing. *arXiv preprint arXiv:2503.06809*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Preprint*, arXiv:2307.06350.

Zhaoning Jin, Ying Fu, Ji Zhou, Wensheng Xie, Linlin Zhang, and Shujian Liu. 2024. Adaptive image quality assessment via teaching large multimodal model to compare. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. 2023. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *Preprint*, arXiv:2310.01506.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023a. Imagic: Text-based real image editing with diffusion models. *Preprint*, arXiv:2210.09276.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023b. Imagic: Text-based real image editing with diffusion models. *Preprint*, arXiv:2210.09276.

Amir Kazerouni, Ehsan Khodapanah Aghdam, Mohammad Heidari, Ramin Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. 2022. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*.

Benjamin Kidder, John Smith, and Alice Lee. 2024. Advanced image generation for cancer using diffusion models. *BioMethods*, 9(1):bpae062.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. *Preprint*, arXiv:2110.02711.

Joni Korhonen and Junyu You. 2012. Peak signal-to-noise ratio for image quality assessment. In *Proceedings of the 2012 International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 231–236.

Su-In Lee, Hyun Kim, Jae Park, and Min Choi. 2024. Monet: Medical concept retriever for transparent ai in dermatology. *Nature Medicine*, 30:123–130.

Han Li, Hao He, Zhiwen Zhang, Qian Liu, Shengyao Ding, Chaofeng Chen, and Weisi Lin. 2024a. M3-agiqa: Multimodal, multi-round, multi-aspect ai-generated image quality assessment. *arXiv preprint arXiv:2502.15167*.

Wei Li, Hao Chen, and Rui Zhang. 2024b. Synthetic data generation by diffusion models. *National Science Review*, 11(8):nwae276.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. 2024a. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. *Preprint*, arXiv:2403.03431.

Qizhe Liu, Weidi Tang, and Bolei Wang. 2024b. A comprehensive study of multimodal large language models for image quality assessment. *arXiv preprint arXiv:2403.10854*.

Kangbo Ma. 2025. Textdiffseg: Text-guided latent diffusion model for 3d medical image segmentation. *arXiv preprint arXiv:2504.11825*.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. 2024. I2ebench: A comprehensive benchmark for instruction-based image editing. *Preprint*, arXiv:2408.14180.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations. *Preprint*, arXiv:2108.01073.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Null-text inversion for editing real images using guided diffusion models. *Preprint*, arXiv:2211.09794.

Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. 2024. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. *Preprint*, arXiv:2412.04301.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *Preprint*, arXiv:2112.10741.

OpenAI. 2024. Gpt-4o: Openai's new multimodal model. Accessed: 2025-03-05.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Lemuel Puglisi, Daniel C. Alexander, and Daniele Ravì. 2024. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. *Preprint*, arXiv:2405.03328.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Hareesh Ravi, Sachin Kelkar, Midhun Harikumar, and Ajinkya Kale. 2023. Preditor: Text guided image editing with diffusion prior. *Preprint*, arXiv:2302.07979.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *Preprint*, arXiv:2401.14159.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Preprint*, arXiv:2205.11487.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

S Taylor, JM Brown, K Gupta, JP Campbell, S Ostmo, RVP Chan, J Dy, D Erdogmus, S Ioannidis, SJ Kim, J Kalpathy-Cramer, MF Chiang, Imaging, and Informatics in Retinopathy of Prematurity Consortium. 2019. Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. *JAMA Ophthalmology*, 137(9):1022–1028.

Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. 2023a. Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path. *Preprint*, arXiv:2303.16765.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson,

and William Chan. 2023b. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *Preprint*, arXiv:2212.06909.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. 2022. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *arXiv preprint arXiv:2212.06909*.

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

Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. 2025. Paint by inpaint: Learning to add image objects by removing them first. *Preprint*, arXiv:2404.18212.

Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. 2023. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. *arXiv preprint arXiv:2301.11798*.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. *Preprint*, arXiv:2012.03308.

Keting Xu, Kaixuan Lin, Siwei Zhang, Zeyu Chen, Wei Hong, Yizhe Wang, and Xiang Ren. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*.

Serin Yang, Hyunmin Hwang, and Jong Chul Ye. 2024. Freestyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*.

Ahmet Burak Yildirim, Vedat Bagdat, Aykut Erdem, and Erkut Erdem. 2023. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03216*.

Li Zhang, Mei Wang, and Qiang Liu. 2024a. Application of artificial intelligence-assisted image diagnosis in medical imaging education. *BMC Medical Education*, 24(1):102.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. *Preprint*, arXiv:2302.05543.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595.

Yiming Zhang, Tao Chen, Jiawei Zhang, Ying Wang, and Weisi Lin. 2024b. Grounding-iqa: Multimodal language grounding model for image quality assessment. *arXiv preprint arXiv:2411.17237*.

Yuxin Zhang, Ming Li, Xiaowei Wang, Yuxin Zhao, and Lei Zhang. 2024c. Text-guided diffusion enhances rare thyroid cancer ai. *Computers in Biology and Medicine*, 168:107728.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. 2025a. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *Preprint*, arXiv:2504.20690.

Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Gorkem Durak, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. 2023b. Diffboost: Enhancing medical image segmentation via text-guided diffusion model. *arXiv preprint arXiv:2310.12868*.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. 2025b. Sine: Single image editing with text-to-image diffusion models. *Preprint*, arXiv:2212.04489.

## A Introduction

Figure 7 illustrates the limitations of controlling text and image guidance scales in diffusion-based image editing. Specifically, for the task of "adding a missing tooth," varying these scales fails to yield a satisfactory result when using the InstructPix2Pix model. While increasing the text guidance scale emphasizes the semantic prompt, and higher image guidance preserves visual fidelity to the original image, neither direction successfully produces the desired anatomical modification. This outcome suggests that simply tuning global guidance weights is insufficient for achieving fine-grained, localized edits in medical or detail-critical domains. As shown in Figure 7, the generated outputs either omit the new tooth entirely or introduce unnatural artifacts, underscoring the need for more controllable and spatially-aware editing approaches.

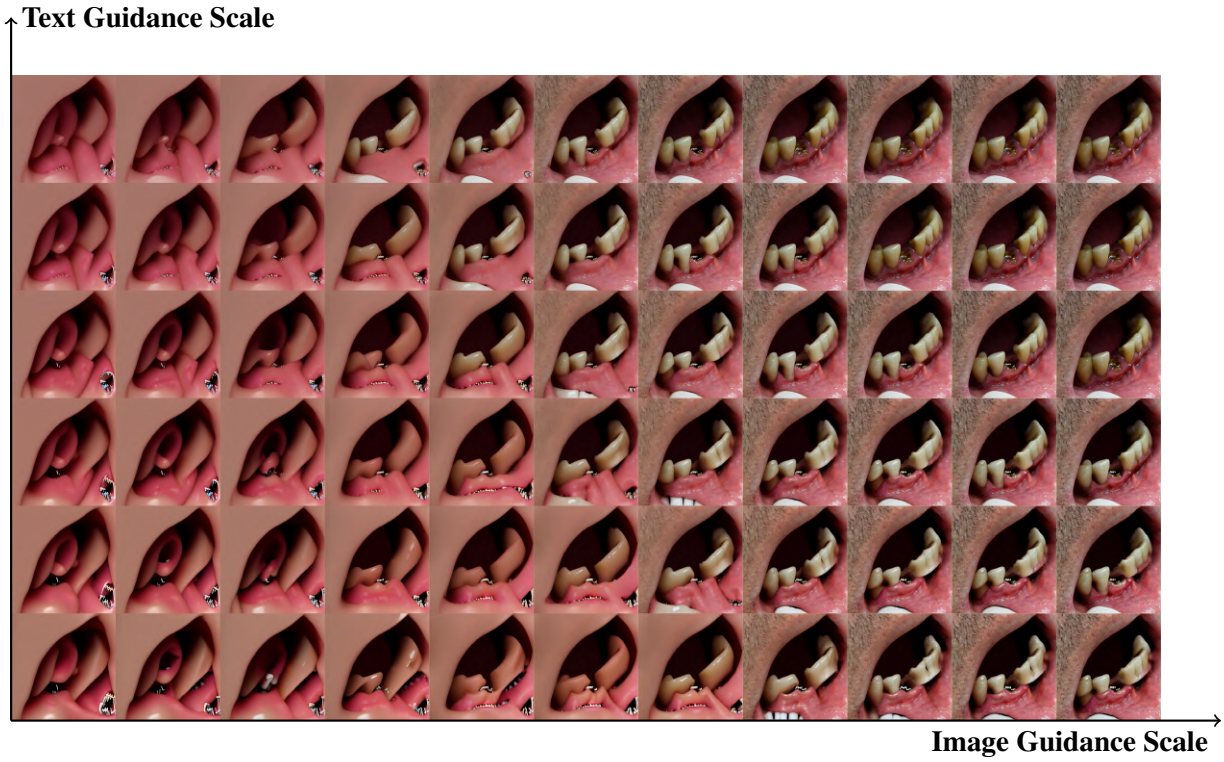



Figure 7: Visualization of the effect of varying text and image guidance scales for the task of adding a missing tooth. Despite adjustments, the desired edit could not be achieved by InstructPix2Pix.

## B Details of MedEBench Dataset

Most images were collected using Google Search with Creative Commons license filters. Some additional images were taken from figures in publicly available open-access papers published under the CC BY 4.0 license, including journals such as MDPI, Frontiers, PLOS, eLife, BMC Bioinformatics, BMC Medicine and Hindawi. To promote transparency, we provide source URLs for all images whenever possible.

Table 5 summarizes the distribution of organ-related tasks included in MedEBench, focusing on tasks with more than five samples.

Each sample in the MedEBench dataset is defined by a structured metadata entry that includes the editing prompt, a description of the expected change, a pair of previous and ground truth images, and an annotated ROI mask on the previous image. The URL of the source image is also provided. Examples are shown in Fig. 8 and Fig. 9.

## C Baseline Models Implementation Detail

We summarize the inference-time configurations for each baseline model as follows:

| Organ | Task | Sample Count |
|---|---|---|
| Ears | Reconstruct ear. | 31 |
| | Reconstruct auricle. | 11 |
| Eyes | Reconstruct missing or injured eye. | 35 |
| | Reconstruct eyelid or lower eyelid skin. | 29 |
| | Create double eyelids. | 29 |
| | Remove eyelid styes (chalazion). | 14 |
| | Correct lower eyelid ectropion. | 10 |
| | Remove conjunctival nevus. | 8 |
| | Remove eye bags. | 8 |
| | Remove eyelid xanthalasma. | 6 |
| Face | Reduce facial wrinkles. | 25 |
| | Remove or revise facial scars. | 21 |
| | Reconstruct cheek or forehead skin. | 17 |
| | Remove facial redness. | 17 |
| | Remove excess fat from neck. | 15 |
| | Lift neck and face by tightening skin. | 10 |
| | Remove facial acne, bumps, or cysts. | 7 |
| Feet | Repair or reconstruct toenail. | 30 |
| Gastrointestinal Tract | Remove intestinal polyps or adenomas. | 15 |
| Hair | Make hair thicker. | 150 |
| | Make beard thicker. | 13 |
| | Make eyebrow thicker. | 8 |
| Hands | Complete missing finger or fingernail. | 18 |
| | Improve hand appearance by injectable filler. | 7 |
| Mouth | Reconstruct damaged lip or lip skin. | 20 |
| | Perform lip augmentation. | 11 |
| Nose | Reconstruct or repair nose skin. | 107 |
| | Remove or fade scar on nose. | 23 |
| | Remove or beautify rhinophyma. | 16 |
| Skin | Remove moles, nevi, or black marks. | 136 |
| | Reconstruct damaged skin or scalp. | 48 |
| | Remove varicose veins. | 36 |
| | Remove brown spots or pigmentation. | 31 |
| | Remove or fade scars. | 20 |
| | Remove black birthmarks to even skin tone. | 14 |
| Spine and Bones | Correct spine alignment. | 10 |
| | Fix fractures with screws or splints. | 6 |
| Teeth | Remove wisdom teeth. | 31 |
| | Remove stains, tartar, or plaque from teeth. | 22 |
| | Implant or add missing teeth. | 13 |
| | Repair or restore damaged teeth. | 13 |
| Torso | Perform body liposuction and skin tightening. | 36 |

Table 5: Tasks with More Than 5 Samples per Organ

```
{
  "Id": 1,
  "Organ": "Teeth",
  "Task": "Implant or add missing teeth.",
  "Prompt": "Add a tooth in the missing area.",
  "Rephrased_prompt": "Place a tooth where one is absent",
  "Detailed_description": "The image modality is intraoral photography; the addition of a tooth in the lower dental arch was performed to fill the gap, resulting in a complete and continuous row of teeth with natural alignment and spacing.",
  "Previous_image": "editing/previous/1.png",
  "GroundTruth_image": "editing/changed/1.png",
  "ROI_mask": "editing/previous_mask/1.png",
  "url": "https://dentistpeshawar.pk/wp-content/uploads/2024/07/extraction-and-implant-same-day.jpg"
}
```


Figure 8: Example metadata entry in MedEBench.

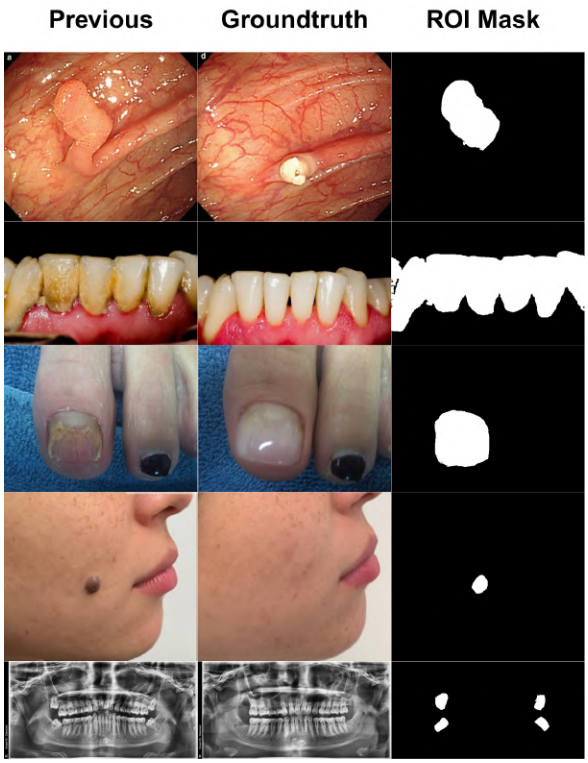



Figure 9: Example groups of previous groundtruth images and ROI mask.

- **InstructPix2Pix**: `DIFFUSION_STEPS` = 50, `image_guidance_scale` $\in \{1.55, 1.6, 1.65\}$, `guidance_scale` $\in \{7.5, 7.6\}$.
- **Imagic**: `num_inference_steps` = 50, `alpha` $\in \{1.3, 1.35\}$, `guidance_scale` $\in \{7.5, 7.6\}$.
- **InstructDiffusion**: `cfg_text` = 5.0, `cfg_image` = 1.25.
- **Paint-by-Inpaint**: `DIFFUSION_STEPS` = 50, `image_guidance_scale` = 1.7, `guidance_scale` = 7.0.
- **ICEdit**: `num_inference_steps` = 28, `guidance_scale` = 50.
- **SEED-X**: `num_inference_steps` = 50, instruction format: `"Edit this image: "` + editing prompt.
- **Gemini 2.0 Flash**: prompt format: `"You are good at image editing. Here is the image editing instruction: "` + editing instruction.

## D Human Evaluation Criteria

To validate the effectiveness of our automated metrics, we conducted a human evaluation study using a relative ranking protocol. Two expert annotators independently assessed a subset of edited images across three dimensions—*Editing Accuracy* (EA), *Contextual Preservation* (CP), and *Visual Quality* (VQ)—which align with the dimensions used in our automated evaluation framework. The evaluation interface and annotator instructions are illustrated in Fig. 10. For each organ $o \in \mathcal{O}$, we randomly sampled 20 editing tasks, resulting in a total of 260 edited image samples. Outputs from multiple models were collected using identical inputs and prompts to ensure fair comparison. Two annotators with bachelor's degrees were hired from a crowdsourcing platform, compensated at a rate of 4.2 HKD per sample. They were instructed to rank the model outputs for each sample along all three dimensions. The inter-annotator agreement, measured by the Spearman rank correlation coefficient, reached 0.91, indicating excellent consistency and high annotation reliability.

## E Automated Evaluation Pipeline

Figure 11 provides an overview of our automatic evaluation pipeline, which consists of two complementary components: GPT-4o-based judgment and masked SSIM measurement.

On the left side of the figure, we illustrate a two-step process for MLLM-based evaluation. In Step 1,

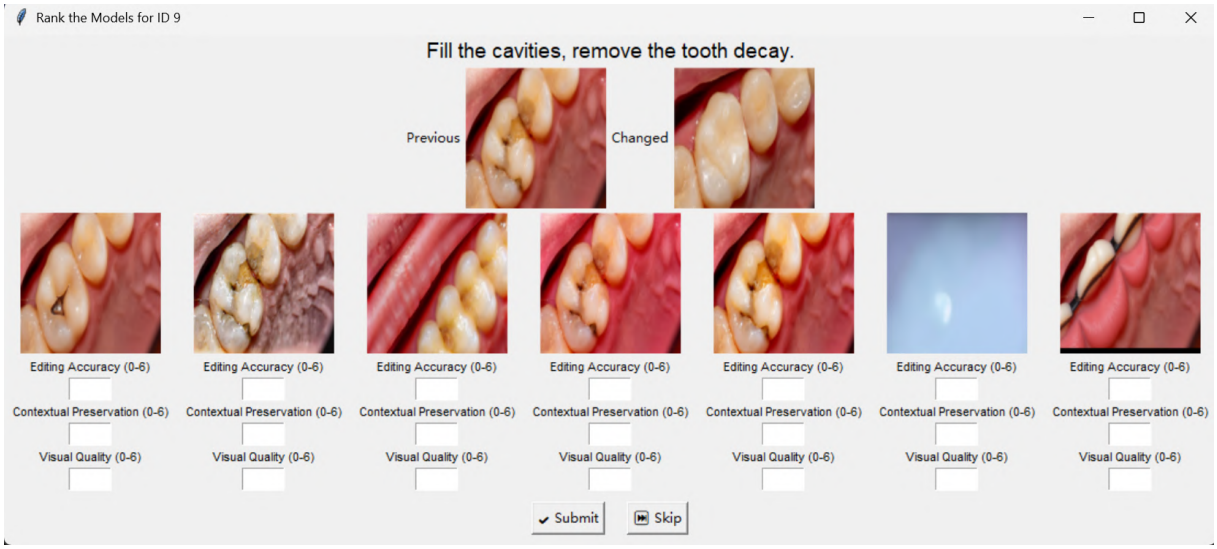



Figure 10: **Human Evaluation Panel.** Annotators were instructed to rank edited images based on three dimensions: **Editing Accuracy (EA)** — whether the intended anatomical modification has been correctly and plausibly applied as described in the prompt; **Contextual Preservation (CP)** — whether unedited regions maintain their original anatomical structure without unintended alterations; and **Visual Quality (VQ)** — the overall perceptual quality of the image, including the seamlessness of edits, absence of artifacts, clarity, and realistic color fidelity.

GPT-4o compare the original image ($I_{\text{prev}}$) and the edited image ($I_{\text{edit}}$) from a TIE model to generate a description of the actual visual difference. Simultaneously, we also use GPT-4o to produce a reference description of the expected change, given the edit prompt and the pair $(I_{\text{prev}}, I_{\text{gt}})$, where $I_{\text{gt}}$ is the ideal target image. In Step 2, GPT-4o compares the actual and expected change descriptions and produces two scores: *Editing Accuracy*, which reflects how well the edit aligns with the prompt, and *Visual Quality*, which assesses the perceptual realism and consistency of the output.

On the right side of the figure, we assess *Contextual Preservation* using masked structural similarity. An ROI mask is generated using a prompt-guided method such as Grounded-SAM or through manual selection. We then compute SSIM between $I_{\text{prev}}$ and $I_{\text{edit}}$, restricted to the masked region's complement. This quantifies how much of the image outside the intended region was unnecessarily altered.

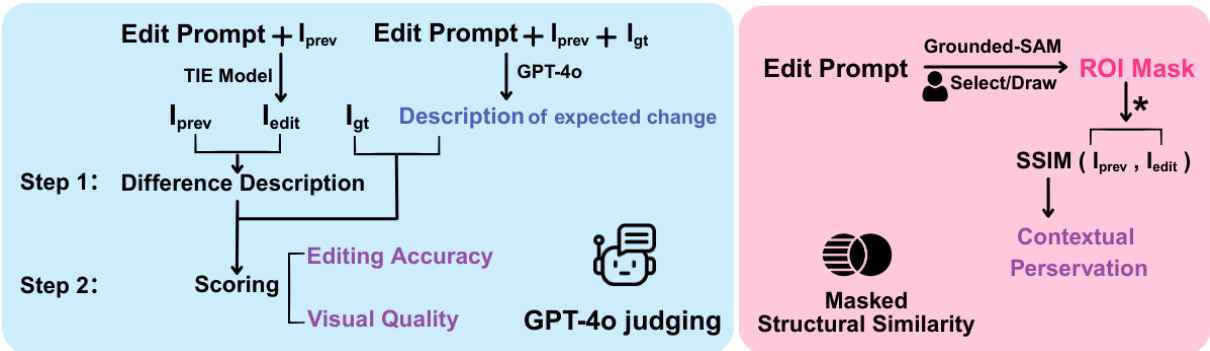



Figure 11: Automatic Evaluation Pipeline in detail.

### E.1 An Example to Demonstrate the Evaluation Process

Figure 12 presents a representative evaluation result for a sample on Contextual Preservation. The corresponding GPT-4o-based evaluation of editing accuracy and visual quality is shown below.

**Edit Prompt:** "Remove the wisdom teeth."

**Description of Expected Change:** "Panoramic dental radiograph: Wisdom teeth in the upper and lower jaws were removed, resulting in a post-surgical appearance with space in the alveolar bone where the teeth previously resided."

Figure 12: Evaluation example for contextual preservation. The contextual preservation score (masked SSIM) is 0.83.

**GPT-4o Evaluation Output for the Sample in Fig. 12**

**Visual Difference Description**

- **Lower Wisdom Tooth (Mandibular Third Molar):**
  - The impacted lower left wisdom tooth remains clearly visible in the edited image.
  - There is no evidence of surgical removal or space in the alveolar bone.
  - No post-surgical bone contour or socket is visible — it is identical to the input.
- **Upper Wisdom Tooth (Maxillary Third Molar):**
  - The upper left third molar is also present and unchanged in the edited image.
  - Again, no extraction space or bone remodeling appears.
- **No Visible Surgical Artifacts:**
  - There are no additions or edits indicating removal, such as bone healing patterns, shadow variations, or softening around the third molar regions.
- **Image Tone and Contrast:**
  - The edited image has slightly more contrast and minor brightness variation.
  - These changes appear incidental and are not related to anatomical edits.

**Step 2: Evaluation**

- **Editing Accuracy: 0/10**
  - Neither the upper nor lower wisdom teeth were removed.
  - The anatomical regions remain untouched.
  - No attempt at post-extraction rendering (e.g., healed socket) is visible.
  - The edit does not align with the prompt at all.
- **Visual Quality: 8/10**
  - The edited image retains good clarity and sharpness.
  - The contrast is slightly increased but does not degrade quality.
  - No artifacts, glitches, or unrealistic elements are introduced — the image remains diagnostically useful.

### E.2 Additional Editing Results

Table 6 provides visual comparisons of editing results from 7 baseline models. Each example shows the input, ground truth, and outputs from various methods, with corresponding scores for EA, CP, and VQ.

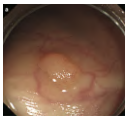
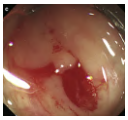
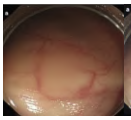
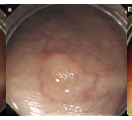
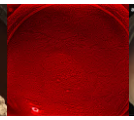
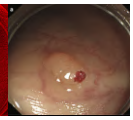
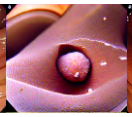
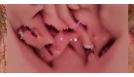
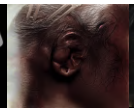
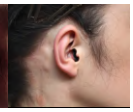
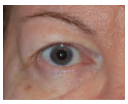
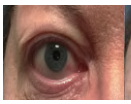
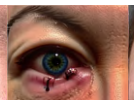
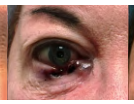
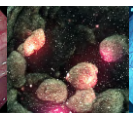
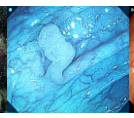
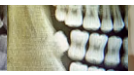
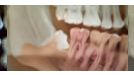
| Previous | Truth | Gemini2 | SeedX | Imagic | IP2P | InstructDiff. | PaintByInpaint | ICEdit |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  | | | |  |  |
| Remove the diminutive polyp. | | 0.8/0.8/0.9 | 0.1/1.0/1.0 | 0.9/0.9/0.9 | 0.6/0.7/1.0 | 0.0/0.3/0.9 | 0.2/0.2/0.8 | 0.9/0.9/1.0 |
| | | | | | | | | |
| Remove the dental black stains. | | 0.9/0.8/0.8 | 0.7/0.5/0.9 | 0.9/0.8/0.7 | 0.5/0.5/0.9 | 0.5/0.7/0.7 | 0.2/0.1/0.6 | 0.0/0.5/0.8 |
| | | |  | | | | | |
| Remove intestinal polyps. | | 0.8/0.8/0.8 | 0.2/0.3/0.2 | 0.2/0.7/0.5 | 0.3/0.7/0.8 | 0.1/0.5/0.7 | 0.0/0.3/0.3 | 0.2/0.6/0.7 |
| | | | | |  | | | |
| Fix damaged front teeth. | | 0.9/0.8/0.9 | 0.0/0.5/0.8 | 0.7/0.6/0.5 | 0.0/0.0/0.7 | 0.2/0.5/0.6 | 0.9/0.9/0.7 | 0.0/0.4/0.9 |
| | | | | | | |  |  |
| Reconstruct the ear. | | 0.9/0.8/0.6 | 0.2/0.4/0.7 | 0.7/0.6/0.4 | 0.2/0.3/0.9 | 0.0/0.3/0.3 | 0.0/0.2/0.4 | 0.8/0.7/0.9 |
| |  |  | |  |  | | | |
| Reconstruct lower eyelid. | | 0.9/0.9/0.6 | 0.4/0.5/0.9 | 0.7/0.9/0.6 | 0.3/0.8/0.9 | 0.2/0.9/0.9 | 0.4/0.8/0.6 | 0.2/0.9/0.7 |
| | | | | | |  |  | |
| Remove intestinal adenoma. | | 0.8/0.6/0.9 | 0.7/0.7/0.7 | 0.2/0.5/0.3 | 0.0/0.6/0.8 | 0.0/0.3/0.2 | 0.0/0.4/0.5 | 0.0/1.0/0.9 |
| | | | |  |  | | | |
| Remove wisdom teeth. | | 0.6/0.7/0.8 | 0.5/0.4/0.8 | 0.7/0.6/0.4 | 0.2/0.1/0.8 | 0.7/0.6/0.5 | 0.0/0.2/0.5 | 0.2/0.6/0.8 |
| | | | | | | | | |
| Remove dental black stains. | | 0.9/0.8/0.9 | 0.0/0.3/0.8 | 0.2/0.5/0.8 | 0.8/0.7/0.9 | 0.0/0.0/0.7 | 0.0/0.2/0.8 | 0.1/0.3/0.9 |

Table 6: Visual comparison of editing results. Each row includes the previous and ground truth images, followed by outputs from seven models. Scores below each output denote EA (Editing Accuracy), VQ (visual quality), and CP (masked SSIM) (all in [0, 1] range).

## F Learning Paradigm Comparison

This section provides a detailed analysis of the two learning paradigms—*fine-tuning* and *in-context learning*—as introduced in Section 3.4. Table 7 reports results across six representative medical image editing tasks, with Editing Accuracy (EA), Contextual Preservation (CP), and Visual Quality (VQ) reported per setting. Fine-tuning varies training samples per task (0, 4, 8, 16), while in-context learning varies the number of demonstrations (0, 1, 2, 3).

**Fine-tuning (InstructPix2Pix).** Fine-tuning shows consistent gains on structured reconstruction tasks such as *Reconstruct Ear* and *Reconstruct Nose*, where EA steadily improves with more training samples. In contrast, performance on removal tasks (*Remove Wisdom Teeth*, *Remove Moles*) often plateaus or declines, suggesting that excessive supervision can lead to overfitting or degraded semantic precision in fine-grained edits.

**In-context Learning (Gemini 2).** Gemini 2 performs well in the zero-shot setting on several tasks, but shows limited benefit from additional demonstrations. In some cases, performance declines with more shots—for example, EA for *Reconstruct Ear* drops from 0.75 to 0.72 from 0-shot to 3-shot. This suggests difficulty in distinguishing test inputs from exemplars, limiting the model's ability to generalize in pixel-level editing tasks.

Overall, fine-tuning is more effective when modest amounts of task-specific supervision are available, particularly for structured and spatially consistent reconstruction tasks. In contrast, in-context learning with large-scale models like Gemini offers reasonable generalization in zero-shot settings but does not scale effectively with additional examples. These findings highlight the need for improved prompting strategies or architectural enhancements to enable reliable few-shot medical image editing via in-context learning.

**(a) InstructPix2Pix (Finetuning Samples)**

| Task | 0s | 4s | 8s | 16s |
|---|---|---|---|---|
| Reconstruct ear | 0.34 / 0.90 / 0.67 | 0.35 / 0.90 / 0.69 | 0.38 / 0.87 / 0.67 | 0.43 / 0.86 / 0.68 |
| Remove wisdom teeth | 0.31 / 0.87 / 0.42 | 0.31 / 0.85 / 0.40 | 0.29 / 0.81 / 0.39 | 0.30 / 0.83 / 0.40 |
| Remove moles | 0.46 / 0.83 / 0.72 | 0.45 / 0.85 / 0.73 | 0.41 / 0.84 / 0.73 | 0.37 / 0.82 / 0.68 |
| Reconstruct nose | 0.68 / 0.88 / 0.77 | 0.70 / 0.86 / 0.74 | 0.75 / 0.86 / 0.75 | 0.77 / 0.87 / 0.75 |
| Remove varicose veins | 0.62 / 0.75 / 0.32 | 0.64 / 0.73 / 0.31 | 0.67 / 0.74 / 0.35 | 0.67 / 0.77 / 0.34 |
| Reconstruct toenails | 0.17 / 0.90 / 0.48 | 0.19 / 0.90 / 0.48 | 0.20 / 0.90 / 0.48 | 0.25 / 0.90 / 0.48 |

**(b) Gemini 2 (In-Context Samples)**

| Task | 0s | 1s | 2s | 3s |
|---|---|---|---|---|
| Reconstruct ear | 0.75 / 0.70 / 0.84 | 0.73 / 0.65 / 0.83 | 0.71 / 0.63 / 0.82 | 0.72 / 0.57 / 0.81 |
| Remove wisdom teeth | 0.70 / 0.81 / 0.79 | 0.72 / 0.78 / 0.80 | 0.70 / 0.77 / 0.79 | 0.70 / 0.73 / 0.77 |
| Remove moles | 0.87 / 0.84 / 0.77 | 0.84 / 0.77 / 0.78 | 0.83 / 0.78 / 0.80 | 0.84 / 0.72 / 0.77 |
| Reconstruct nose | 0.65 / 0.85 / 0.76 | 0.67 / 0.79 / 0.75 | 0.65 / 0.78 / 0.76 | 0.65 / 0.80 / 0.75 |
| Remove varicose veins | 0.73 / 0.77 / 0.63 | 0.73 / 0.76 / 0.65 | 0.74 / 0.75 / 0.66 | 0.73 / 0.75 / 0.67 |
| Reconstruct toenails | 0.81 / 0.83 / 0.80 | 0.82 / 0.80 / 0.82 | 0.84 / 0.82 / 0.81 | 0.82 / 0.80 / 0.80 |

Table 7: Evaluation results of (a) Finetuned InstructPix2Pix with 0, 4, 8, and 16 finetuning samples per task, and (b) Gemini 2 in-context learning with 0, 1, 2, and 3 in-context samples. Each cell reports Editing Accuracy (EA), Contextual Preservation (CP), and Visual Quality (VQ) in the format: EA / CP / VQ.

## G Attention Grounding

Figure 13 illustrates the detailed average Intersection over Union (IOU) scores across tasks and anatomical regions.

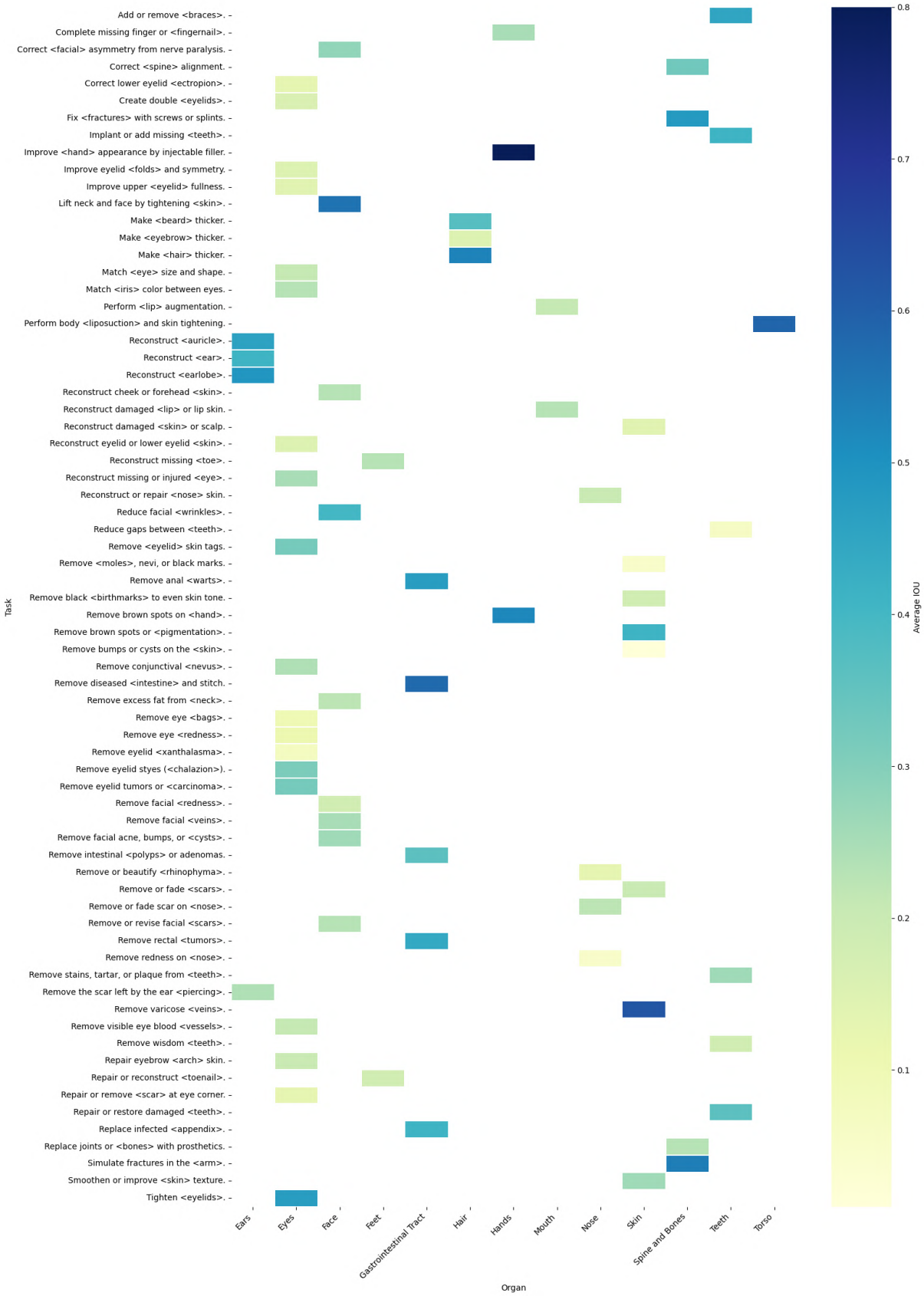



Figure 13: Average IOU scores for different tasks across anatomical regions. Concepts are annotated in angle brackets.

# H Prompt Templates

## H.1 GPT-4o Prompt for Generating Change Descriptions

**Input:** A concatenated image (previous on the left, changed on the right) and an editing instruction.

**GPT-4o Change Description Prompt**

You are a medical expert. Your task is to use your extensive medical knowledge to describe the visual change from the left image to the right image, guided by the clues provided in the editing instruction. Based on these inputs, generate a detailed, precise, and clinically relevant description of the visual change that was applied.
Your response must clearly state:

- The image modality
- The action performed and its target entity
- The expected appearance or anatomical result after the change

Format your output as a single formal instruction sentence describing how the right image was derived from the left.

**Instruction:** "<editing_instruction>"

**Concatenated image:** <base64-encoded image with previous on the left and changed on the right>

## H.2 GPT-4o Judging Prompt (with editing prompt)

**Input:** Previous image, ground truth image, edited image, and editing prompt.

**GPT-4o Judging Prompt**

I have an image editing task. Here's the editing prompt:
"<editing_prompt>"

Here is the input image:
<base64-encoded previous image>

Here is the ground truth image:
<base64-encoded ground truth image>

Here is the edited image:
<base64-encoded edited image>

You are good at evaluating image editing. Evaluate the edit using the following steps:

**Step 1: Visual Difference Description**
Compare the input and edited images. Describe all visible differences between them, including:

- Additions, removals, or modifications of visual elements.
- Emphasize the extent of the changes and specify which anatomical regions were affected.

List the differences clearly, item by item.

**Step 2: Evaluation (Three Scores)**
Provide the following scores with brief justifications:

- **1. Editing Accuracy (0–10):**
  Score strictly based on alignment with the editing prompt. You may use the ground truth image as a reference, though it may not be pixel-aligned. Deduct points for inaccuracies or missing elements.
- **2. Visual Quality (0–10):**
  Compare the overall visual quality of the edited image to the original. Consider clarity, sharpness, blur, artifacts, realism, and consistency.

**Please respond in the following format:**

- **Visual Difference Description:**
  - 1. ...
  - 2. ...
  - ...
- **Editing Accuracy:** [Score]/10, [Reasons]
- **Visual Quality:** [Score]/10, [Reasons]

### H.3 GPT-4o Judging Prompt (with change description)

**Input:** Previous image, ground truth image, edited image, and change description.

**GPT-4o Judging Prompt**

I have an image editing task. Here's the description of the expected change:
`"<change_description>"`

Here is the input image:
`<base64-encoded previous image>`

Here is the ground truth image:
`<base64-encoded ground truth image>`

Here is the edited image:
`<base64-encoded edited image>`

You are good at evaluating image editing. Evaluate the edit using the following steps:

**Step 1: Visual Difference Description**
Compare the input and edited images. Describe all visible differences between them, including:

- Additions, removals, or modifications of visual elements.
- Emphasize the extent of the changes and specify which anatomical regions were affected.

List the differences clearly, item by item.

**Step 2: Evaluation (Three Scores)**
Provide the following scores with brief justifications:

- **1. Editing Accuracy (0–10):**
  Score strictly based on alignment with the editing prompt. You may use the ground truth image as a reference, though it may not be pixel-aligned. Deduct points for inaccuracies or missing elements.
- **2. Visual Quality (0–10):**
  Compare the overall visual quality of the edited image to the original. Consider clarity, sharpness, blur, artifacts, realism, and consistency.

**Please respond in the following format:**

- **Visual Difference Description:**
  - 1. ...
  - 2. ...
  - ...
- **Editing Accuracy:** [Score]/10, [Reasons]
- **Visual Quality:** [Score]/10, [Reasons]