BEAR-MED: FROM A CLINICALLY GROUNDED BENCHMARK TO EXPERT-ALIGNED REASONING IN MEDICAL IMAGE FORENSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical image forensics extends beyond binary detection: systems must distinguish local edits from global synthesis, localize affected regions, and justify verdicts with observable evidence. Existing benchmarks remain limited in scale, imaging diversity, forgery coverage, clinical plausibility, or spatial fidelity, while current methods often separate detection from localization, append post-hoc explanations, or derive chain-of-thought supervision from general-purpose multimodal models without explicit forensic evidence. We introduce BEAR-Med-1M, a million-image benchmark spanning four clinical imaging categories and five local-to-global mechanisms: copy-move, splicing, lesion insertion, lesion removal, and full-image synthesis. It targets medically meaningful, anatomically plausible regions and uses bidirectional restoration for lesion insertion and removal, producing spatially faithful local forgeries with low out-of-box spillover. We also introduce BEAR-Med, an expert-aligned model whose four frozen experts capture local traces, image-wide consistency, generation texture, and medical normality. Their evidence is distilled offline into expert-guided CoT targets, while inference uses only the original image. BEAR-Med generates a pre-hoc analyze–localize–decide sequence, placing evidence-based reasoning before localization and verdict rather than explaining predictions post hoc. A segmented objective separately balances reasoning, box, and verdict tokens, followed by GRPO with verifiable task and spatial rewards. BEAR-Med leads dense-localization and multimodal reasoning baselines in accuracy, F1, all reported localization metrics, and authentic-image error control, while retaining strong cross-dataset detection and localization.

1 INTRODUCTION

Manipulated clinical images can duplicate or transplant a finding, insert or erase a lesion, or fabricate an entire examination. Such content can contaminate clinical records or medical AI datasets and falsify evidence in publications or insurance claims (Zingarini et al., 2024; Li et al., 2025; Chen et al., 2026). In these high-stakes settings, a forensic system must move beyond binary authenticity: it should identify the manipulation scope, localize affected regions for local edits, abstain from local boxes for authentic or globally synthesized images, and justify its verdict with observable, clinically interpretable evidence.

Existing medical forgery benchmarks remain limited in scale, imaging categories, and manipulation mechanisms (Zingarini et al., 2024; Li et al., 2025; Chen et al., 2026). Yet broader coverage alone does not ensure clinically meaningful local forgeries. Lesion insertion and removal must target anatomically compatible sites; arbitrary placement creates implausible, easily detected artifacts unlike deliberate manipulation. Spatial fidelity matters as well: inpainting may introduce meaningful visual changes beyond the requested region, making the spatial annotation unreliable. Thus, a high-quality benchmark must control both where an edit is applied and where its visible effect occurs.

General synthetic-image detectors classify authenticity without supporting regions (Wang et al., 2023b; Ojha et al., 2023), while dense localizers provide manipulation maps without explaining forensic mechanisms or clinical plausibility (Guillaro et al., 2023; Zhu et al., 2025). Multimodal detectors add rationales, but these may be post hoc or learned from general-purpose CoT without

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.

explicit forensic evidence (Wen et al., 2025; Chen et al., 2026; Zhu et al., 2026). This leaves a central question: can specialized forensic evidence shape a unified reasoning–localization–decision policy during training without becoming a deployment-time dependency?

We address the data side with BEAR-MED-1M, a benchmark of 500K authentic and 500K forged images spanning planar radiography, CT/MRI, ultrasound, and optical clinical imaging. Its five mechanisms form a local-to-global spectrum: copy-move, splicing, lesion insertion, lesion removal, and full-image synthesis. Lesion insertion and removal use opposite restoration directions at medically meaningful, anatomically plausible regions: one model learns to reconstruct masked lesions and is then used to insert them into healthy regions, while the other learns to reconstruct masked healthy tissue and is then used to replace lesions with healthy tissue. A full paired-image audit tests whether meaningful structural changes remain outside each local box, confirming low out-of-box spillover and that the annotations track the realized edit rather than only the requested generation region.

We address the model side with BEAR-MED, which uses four complementary frozen experts as privileged supervision available only during offline target construction. The local-trace and image-consistency experts expose manipulation artifacts, the generation-texture expert targets synthesis signatures, and the medical-normality expert assesses anatomical and clinical plausibility. A teacher receives the raw image and these expert views and distills their evidence into expert-guided chain-of-thought targets. Using these targets, we train a Qwen3-VL policy (Bai et al., 2025) that receives only the raw clinical image during both training and deployment. The policy autoregressively emits `[REASONING]`, `[BBOX]`, and `[VERDICT]`; this pre-hoc analyze–localize–decide order places evidence before localization and verdict rather than explaining a completed prediction.

Training combines segmented SFT with Grounded GRPO (G-GRPO). Segmented SFT prevents long rationales from diluting box and verdict supervision. G-GRPO extends GRPO (Shao et al., 2024) with verifiable task and spatial rewards. On the BEAR-MED-1M test split, BEAR-MED reaches .995 F1, .839 Box IoU@0.75, and .001 false-positive and false-box rates, leading all evaluated dense-localization and multimodal baselines in accuracy, F1, and every reported localization metric. Expert subgroup and fusion analyses confirm complementary coverage, while ablations isolate expert-guided CoT and segmented SFT. Blinded judges rate its rationales highest across six dimensions; degradation and cross-dataset tests assess robustness beyond the benchmark. Our contributions are:

- BEAR-MED-1M, a clinically grounded million-image benchmark covering four clinical imaging categories and five local-to-global forgery mechanisms, with medically meaningful edits and independently audited spatial fidelity;
- BEAR-MED, an expert-aligned model that distills four complementary frozen experts into raw-image, pre-hoc analyze–localize–decide behavior, together with segmented SFT and verifiable spatial policy optimization;
- A systematic evaluation of detection, strict localization, authentic- image error control, rationale quality, degradation robustness, and cross-dataset transfer against dense and multimodal baselines.

2 RELATED WORK

Clinical image-forensics benchmarks. Clinical benchmarks cover distinct regimes. BioFors covers duplication and sharp transitions in scientific microscopy, blot/gel, FACS, and macroscopy (Sabir et al., 2021). M3Dsynth edits pulmonary nodules in 3D CT (Zingarini et al., 2024); MedForensics covers full synthesis across six modalities and twelve generators (Li et al., 2025); and MedForge-90K provides lesion editing, reasoning, and edit locations in chest radiography, brain MRI, and fundus images (Chen et al., 2026). BEAR-MED-1M combines copy-move, splicing, bidirectional lesion editing, and full synthesis across planar radiography, CT/MRI, ultrasound, and optical clinical imaging. Local edits target medically plausible sites, with paired-image audits checking out-of-box changes in inpainting outputs.

Forgery detection, localization, and reasoning. Generalizable forensic detectors exploit acquisition traces, upsampling statistics, diffusion reconstruction error, and transferable representations (Cozzolino & Verdoliva, 2020; Tan et al., 2024; Wang et al., 2023b; Ojha et al., 2023; Yan et al., 2025). TruFor, PSCC-Net, IML-ViT, SparseViT, and Mesorch localize edits through noise–semantic fusion, spatial–channel correlation, non-semantic transformer features, or multi-scale features (Guillaro et al., 2023; Liu et al., 2022; Ma et al., 2023; Su et al., 2025; Zhu et al., 2025). MemAU-Net and

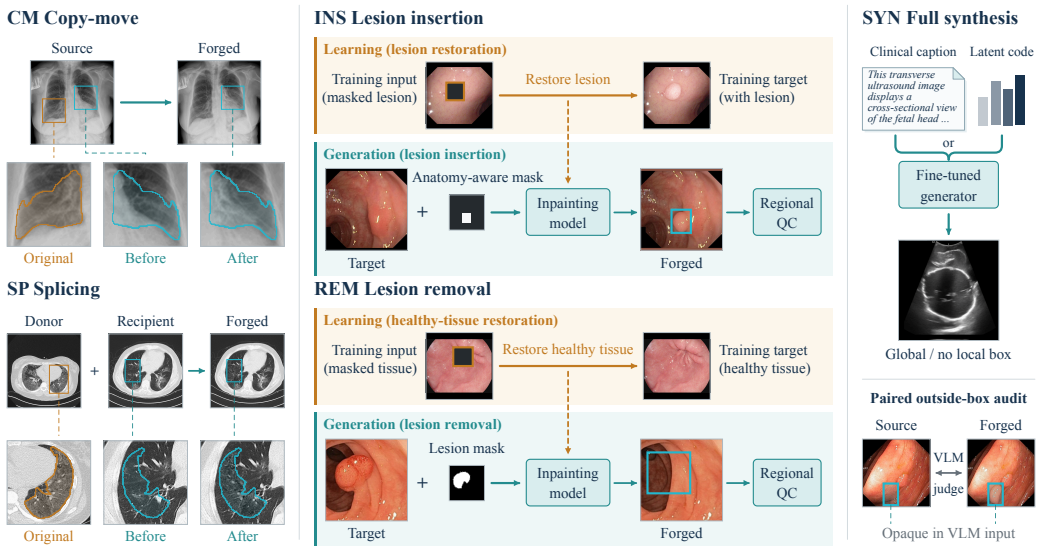


Figure 1: Construction of BEAR-MED-1M. CM and SP transfer findings; INS and REM use bidirectional restoration with regional QC; SYN generates full images without local boxes. Paired source–forgery auditing checks for changes outside the annotated box.

APFL adapt localization to medical images (Wang et al., 2023a; 2024). Multimodal forensic models predict masks with explanations (FakeShield) (Xu et al., 2025), learn fine-grained artifact clues (FakeVLM) (Wen et al., 2025), or unify detection, localization, and explanation (SIDA) (Huang et al., 2025). FakeVLM-R1 combines physical-law counter-evidence with GRPO to reduce authentic-image rejection (Zhu et al., 2026). MedForge obtains rationales from a general-purpose MLLM using doctor-defined guidelines and gold edit locations, and applies policy optimization to a box–reasoning–verdict model (Chen et al., 2026). BEAR-MED instead distills four frozen experts' evidence into a raw-image policy that reasons before localization and verdict.

3 BEAR-MED-1M BENCHMARK

3.1 SCOPE AND SPATIAL TARGETS

BEAR-MED-1M contains 500K authentic and 500K forged images across planar radiography (C1), CT/MRI (C2), ultrasound (C3), and optical clinical imaging, including endoscopy, dermoscopy, and fundus (C4). Each of the five forgery families contributes 100K images, with 25K in each imaging category. Copy-move (CM), splicing (SP), lesion insertion (INS), and lesion removal (REM) have local bounding boxes; full-image synthesis (SYN) has a forged label but no unique local box.

3.2 CLINICALLY CONSTRAINED FORGERY CONSTRUCTION

Source images provide lesion masks or boxes, organ annotations where available, and modality metadata. We curate lesion-to-organ mappings to specify eligible recipient organs and clinical affinity groups to restrict SP to compatible image sources. Figure 1 shows the five mechanisms; the supplement details image sources and annotation preparation.

For CM and SP, we extract a lesion patch for transfer within the same image or to another image, respectively. SP selects a recipient from the same source dataset with the required organ annotation, or from its clinical affinity group. After geometric transformation, the mask must fit the mapped organ or valid tissue. Poisson cloning blends the patch, whose pasted mask defines the output box.

For INS and REM, we construct medically grounded counterfactual forgeries through two inverse restoration tasks rather than asking a generic image editor to add or erase pathology from an instruction alone. The INS model first learns to reconstruct real lesions from masked lesion-bearing images; at

generation, it inserts a lesion into healthy tissue using mask geometry sampled from observed lesions and anatomy-aware placement where organ annotations are available. The REM model learns to reconstruct masked healthy tissue, then replaces a real lesion within its annotated box. Thus, insertion draws on observed lesion morphology, whereas removal draws on learned normal-tissue appearance, with each edit paired to its authentic source. Both directions use FLUX.1-Fill, SDXL-Inpainting, and an SD3/MM-DiT inpainting retrofit (Black Forest Labs, 2024; Podell et al., 2024; Esser et al., 2024). The generation mask defines a local box; a multimodal check accepts INS only when an abnormality is visible in that region and REM only when the targeted lesion is no longer visible.

To audit spillover that can compromise localization supervision, we mask the GT box in source–forgery pairs and use Qwen3.6-27B to flag meaningful changes elsewhere, ignoring exposure and white-balance shifts. It flags 0.64% of 200K INS+REM pairs in BEAR-MED-1M, versus 59.29% of 20K MedForge-90K test pairs (Chen et al., 2026). Two reviewers with medical backgrounds independently assess a blinded, stratified sample of 240 pairs per dataset, yielding mean sampling-weighted rates of 3.40% versus 74.64%. Protocols and full results are in the supplement.

For SYN, a vision-language model captions authentic images to create category-specific image–caption pairs. We fine-tune the text-conditioned FLUX.1-dev, SDXL, SD3.5, and Infinity models on these pairs, then use sampled captions to generate images. The text-free StyleGAN3 model instead learns from authentic images and samples latent codes (Karras et al., 2021; Han et al., 2025).

3.3 CANONICALIZATION AND SOURCE-DISJOINT SPLITS

Authentic and forged images follow the same decode, resize, and lossless encoding path to reduce source-format shortcuts. Images are mapped to a 512×512 canvas and stored as 8-bit RGB. Source images and all their derivatives remain in one split, with patient, study, and volume identifiers grouped when available. This yields 699,771 training, 100,190 validation, and 200,039 test images. Each record stores its label, imaging category, forgery and generator families, and applicable local target.

4 METHOD

4.1 OVERVIEW AND DESIGN PRINCIPLES

BEAR-MED uses four frozen experts to construct privileged supervision (Vapnik & Vashist, 2009) for a raw-image policy (Figure 2). For a clinical image X, their outputs are $Z(X) = \{E_k(X;\phi_k)\}_{k=1}^{4}$, where E_k and ϕ_k denote the k-th expert and its fixed parameters. A teacher combines X, $Z(X)$, clinical-forensic guidance, and private label and box to construct an ordered reasoning–box–verdict target. The student learns to generate the same sequence from the raw image and a fixed instruction.

These complementary priors guide evidence discovery, localization, and decision. The teacher integrates expert strengths, spatial agreement, and conflicts into rationales, making intermediate tasks explicit and helping the student learn the experts' joint regularities.

To avoid the cost of four online expert pipelines and evidence aggregation, we confine expert guidance to training. For fixed ϕ_k, $E_k(X;\phi_k)$ is a deterministic transformation of X, so the data-processing inequality gives

$$\mathrm{I}(Y; E_k(X;\phi_k) \mid \phi_k) \le \mathrm{I}(Y; X \mid \phi_k)\,, \tag{1}$$

where Y is the authenticity label and $\mathrm{I}(\cdot;\cdot)$ denotes mutual information. The sample-dependent evidence exposed by an expert is therefore already encoded in the pixels of X. What the expert contributes is the distribution-level prior encoded by ϕ_k, which organizes those pixels into task-relevant forensic or clinical evidence. We use rationale distillation (Hsieh et al., 2023) to transfer this prior-guided organization into the policy parameters through expert-guided targets, training the policy to produce structured outputs from X alone. The inequality motivates training-only guidance but does not guarantee that a finite student can reproduce expert behavior; Table 6 tests this transfer.

4.2 COMPLEMENTARY FROZEN EVIDENCE

Each expert produces a spatially aligned absolute response map $M_k = E_k(X;\phi_k) \in [0,1]^{32\times32}$, for $k \in \{\mathrm{trace}, \mathrm{cons}, \mathrm{gen}, \mathrm{med}\}$.

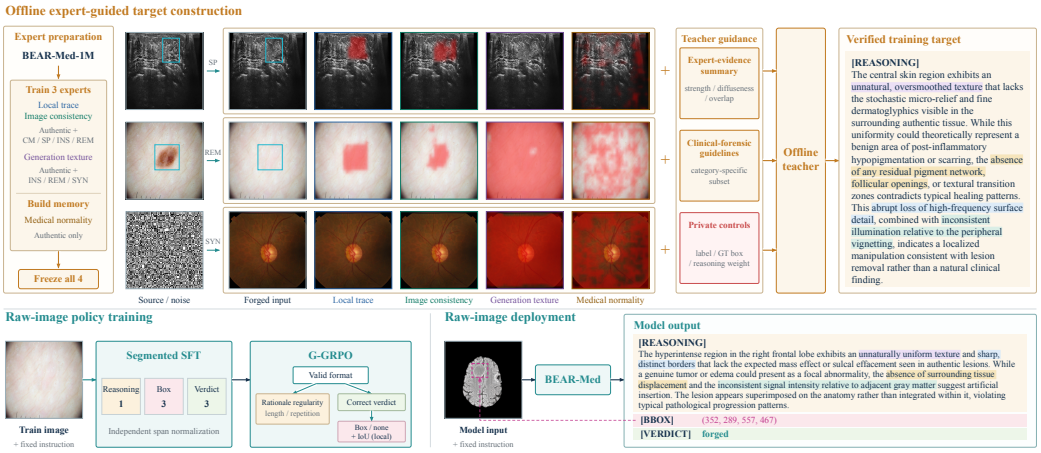


Figure 2: Four frozen experts guide offline rationale construction. The student learns the reasoning–box–verdict sequence from the raw image and a fixed instruction. Colored highlights denote visual cues associated with the corresponding expert priors.

The local-trace expert detects immediate manipulation artifacts. Its multi-view encoder–decoder, inspired by MVSS-Net (Chen et al., 2021), combines the image with median residuals, Sobel boundaries, and distant self-similarity. These views expose duplicated structures, pasted boundaries, blending, and inpainting transitions, providing fine-grained evidence for localization.

The image-consistency expert tests whether each patch follows the low-level acquisition regularities of distant image context. Fine 32×32 queries attend to coarser 16×16 contextual tokens constructed from grayscale, residual, boundary, local-deviation, and chromatic views. This nonlocal comparison reveals inconsistencies that remain inconspicuous within the edited neighborhood.

The generation-texture expert characterizes the image-formation process by combining neighboring-pixel relationships from NPR (Tan et al., 2024) with DIRE diffusion reconstruction error (Wang et al., 2023b). It targets generative texture rather than anatomy and responds strongly to diffusion-based INS, REM, and SYN images, although its evidence need not specify a precise local box.

The medical-normality expert measures deviation from authentic clinical appearance. It compares DINOv2 patch features (Oquab et al., 2023) with PatchCore memories (Roth et al., 2022) built from authentic images and retains the closest match. This prior highlights clinically unusual regions, but a strong response may also reflect uncommon authentic pathology rather than manipulation.

Local trace and image consistency are trained with box-derived patch supervision on authentic and local-forgery images; generation texture uses image-level supervision on authentic, INS, REM, and SYN images. Medical-normality memories use authentic images only. All experts are frozen before CoT construction. Their architectures and training objectives are detailed in the supplement.

4.3 Expert-Guided CoT Construction

Expert response maps are rendered over X to link their patterns to anatomy. An expert-evidence summary records each map's absolute strength, diffuseness, connected high-response regions, and whether experts highlight overlapping regions.

Private controls combine this summary with annotation overlap to set rationale supervision and reasoning weights. For local forgeries, trace and consistency responses that co-localize within the annotated box support stronger supervision, whereas weak, conflicting, or off-box responses reduce it. For SYN, a diffuse generation-texture supports global reasoning, while localized responses remain auxiliary. For authentic images, co-localized strong responses trigger caution.

To supply clinical semantics for interpreting these patterns, we extend MedForge's guideline-grounded inspection (Chen et al., 2026) with category-specific checks of image physics and texture, anatomy, and clinical plausibility. Given X, the rendered expert panel, expert-evidence summary, guideline

subset, and private label/box controls, the teacher writes a rationale while considering an authentic alternative. A verifier rejects references to experts, private supervision, or generation metadata. A deterministic assembler appends the annotated box or `none` and the verdict. The supplement details guideline sources, evidence fields, supervision rules, teacher prompts, and verification.

4.4 Raw-Image Student with Localization-Aware Segmented SFT

Given the raw image X, the student autoregressively produces `[REASONING]` y_r, `[BBOX]` y_b, and `[VERDICT]` y_v. For local forgeries, y_b is a normalized $[x_1, y_1, x_2, y_2]$ box; otherwise it is `none`. The verdict y_v is `authentic` or `forged`. This ordering implements analyze–localize–decide by placing the rationale before the spatial and binary predictions.

Token-averaged cross-entropy lets long rationales dilute box and verdict supervision. SCALe-SFT addresses the analogous reasoning–answer imbalance through independent normalization (Perek et al., 2026). We normalize reasoning, box, and verdict spans within one autoregressive sequence:

$$\bar{\ell}_s = -\frac{1}{|T_s|} \sum_{t \in T_s} \log p_\theta(y_t \mid X, y_{<t}),\ s \in \{r, b, v\}; \qquad \mathcal{L}_{\mathrm{SFT}} = w_r \bar{\ell}_r + 3\bar{\ell}_b + 3\bar{\ell}_v. \tag{2}$$

Here, T_s contains the tokens in span s, and $w_r \in [0, 1]$ is the target-specific rationale weight. Independent normalization removes span-length bias, while the box and verdict weights protect localization and classification without a task-specific head. LoRA (Hu et al., 2022) updates linear modules in the language model, ViT, and aligner/projector; the pretrained weights remain frozen.

4.5 Grounded GRPO

To optimize verdict correctness and box overlap beyond imitation, Grounded GRPO (G-GRPO) samples multiple completions per raw-image prompt from the selected SFT policy and applies group-relative optimization (Shao et al., 2024). We use the hierarchical reward

$$R = 0.2 g_f + g_f \big\{ 0.2 R_{\mathrm{rat}} + g_c \big[1 + 0.5 R_{\mathrm{scope}} + g_l R_{\mathrm{box}} \big] \big\}. \tag{3}$$

Here, $g_f = R_{\mathrm{fmt}}$ and $g_c = R_{\mathrm{cls}}$ indicate valid three-span format and a correct verdict; g_l indicates a local-box target. R_{rat} checks calibrated length and nonrepetition, not clinical correctness. R_{scope} requires a box for local forgeries and `none` for authentic or SYN images, while R_{box} is GT box IoU. Format validity gates all task rewards, and verdict correctness gates scope and localization.

To avoid zero advantages from uniformly rewarded completions, G-GRPO prioritizes train-only frontier prompts with variable SFT-rollout rewards. The supplement details frontier construction, optimization, safeguards, and checkpoint selection.

5 Experiments

5.1 Setup

All trainable methods share the group-disjoint split and validation-only checkpoint selection. Dense localizers train on authentic and local-forgery images, while detection-capable multimodal models also use SYN. Table 1 compares five dense localizers and three task-adapted multimodal baselines using their released objectives and validation criteria. For adaptation on BEAR-Med-1M, each multimodal baseline follows its original CoT-construction protocol, while BEAR-Med uses CoT distilled from the four frozen experts. BEAR-Med trains Qwen3-VL-8B (Bai et al., 2025) with LoRA over the language model, ViT, and aligner/projector. We also evaluate GPT-5.6 Sol, Gemini 3.1 Pro Preview, and Qwen3.8-Max zero-shot. Their APIs provide no calibrated forgery probabilities, so AUC and AP are unavailable.

Detection metrics are accuracy, AUC, AP, precision, recall, F1, and authentic-image FPR. Balanced accuracy equals accuracy on the balanced test split. Dense-model scores use the released classification head or maximum heatmap response. For BEAR-Med, discrete detection metrics use generated verdicts, whereas AUC/AP use normalized label-token probabilities from a separate binary-only query. Localization reports mean Box IoU and Box IoU@0.5/0.75 on local forgeries, plus FBR on authentic images. Threshold searches, decision rules, parsing, and optimization settings are in the supplement.

Table 1: Aggregate detection and localization on the BEAR-MED-1M test set.

	Detection							Localization			
Method	Acc↑	AUC↑	AP↑	Prec.↑	Rec.↑	F1↑	FPR↓	mIoU↑	B@.5↑	B@.75↑	FBR↓
Dense localization baselines											
PSCC-Net	.711	.730	.796	.923	.460	.614	.038	.220	.220	.100	.178
SparseViT	.886	.953	.960	.916	.850	.882	.078	.618	.727	.499	.030
IML-ViT	.922	.971	.975	.947	.893	.919	.050	.677	.800	.555	.029
TruFor	.846	.939	.948	.828	.874	.850	.182	.676	.823	.512	.485
Mesorch	.946	.984	.988	.961	.930	.945	.037	.821	.935	.790	.008
Task-adapted multimodal baselines											
FakeVLM	.787	.793	.724	.732	.905	.809	.332	–	–	–	–
SIDA	.960	**.994**	**.995**	.976	.942	.959	.023	.416	.476	.196	.004
MedForge	.992	.992	.993	.997	.987	.992	.003	.810	.922	.777	.003
Zero-shot closed multimodal models											
GPT-5.6 Sol	.528	–	–	.676	.110	.190	.053	.047	.054	.027	.038
Gemini 3.1 Pro Preview	.598	–	–	.791	.266	.398	.070	.091	.104	.074	.040
Qwen3.8-Max	.524	–	–	.549	.447	.493	.368	.140	.146	.093	.359
BEAR-MED	**.995**	.987	.988	**.999**	**.991**	**.995**	**.001**	**.848**	**.948**	**.839**	**.001**

Table 2: Subgroup detection F1 and localization B@.75 at the aggregate operating points. C1–C4 denote radiography, CT/MRI, ultrasound, and optical imaging.

	Detection F1									Localization B@.75							
	Forgery mechanism					Imaging category				Forgery mechanism				Imaging category			
Method	CM	SP	INS	REM	SYN	C1	C2	C3	C4	CM	SP	INS	REM	C1	C2	C3	C4
PSCC-Net	.554	.609	.587	.641	.362	.118	.587	.669	.881	.137	.227	.026	.008	.003	.141	.130	.123
SparseViT	.778	.802	.817	.825	.528	.919	.728	.920	.958	.474	.560	.474	.490	.498	.202	.514	.785
IML-ViT	.847	.868	.882	.886	.656	.955	.820	.944	.961	.531	.643	.518	.530	.487	.332	.613	.790
TruFor	.639	.666	.671	.675	.470	.893	.674	.912	.968	.456	.576	.510	.508	.464	.350	.551	.685
Mesorch	.906	.910	.914	.914	.724	.961	.879	.962	.982	.739	.803	.787	.829	.806	.589	.803	.961
FakeVLM	.474	.481	.535	.520	.530	.798	.619	.936	.971	–	–	–	–	–	–	–	–
SIDA	.875	.866	.945	.945	.943	.959	.872	1.000	**1.000**	.199	.237	.194	.152	.129	.057	.234	.362
MedForge	.970	.983	.992	.992	.992	**1.000**	.968	**1.000**	**1.000**	.731	.817	.773	.786	.777	.609	.788	.933
BEAR-MED	**.983**	**.994**	**1.000**	**.999**	**1.000**	**1.000**	**.980**	**1.000**	.999	**.799**	**.867**	**.833**	**.856**	**.833**	**.716**	**.844**	**.964**

5.2 MAIN RESULTS ON BEAR-MED-1M

We evaluate in-distribution detection and localization on the complete BEAR-MED-1M test set. All open-source baselines are fine-tuned from their released checkpoints. In Table 1, BEAR-MED achieves .995 F1, .839 B@.75, and .001 FPR/FBR, leading in accuracy, F1, and all four localization metrics. SIDA retains the best AUC/AP (.994/.995) but only .196 B@.75; the strongest comparison localizers, Mesorch and MedForge, reach .790 and .777. Closed models show near-chance zero-shot detection and negligible localization (accuracy .524–.598; B@.75 .027–.093).

Table 2 examines subgroup consistency. Dense localizers are weakest on SYN detection (F1 .362–.724), and C2 is hardest for four of five dense models. BEAR-MED maintains F1 of .983–1.000 across mechanisms and .980–1.000 across categories. It also leads all eight localization subgroups (B@.75 .716 on C2 to .964 on C4). Across all 20,000 SYN images, BEAR-MED emits a spurious box in 0.08% of cases (16 images), irrespective of the predicted verdict.

5.3 EXTERNAL GENERALIZATION TO MEDFORGE-90K

We test cross-dataset transfer after BEAR-MED-1M fine-tuning on all 30,000 MedForge-90K test images (Chen et al., 2026). Detection uses the full split; ALL BOX measures localization on forged images, and CONTAINED retains those whose paired-image audit finds no meaningful change outside the GT box. Table 3 reports adapted results and their changes from the corresponding released checkpoints; the official MedForge model is a target-supervised reference. BEAR-MED leads adapted baselines in BAcc (.769) and macro-F1 (.720), while SparseViT and SIDA retain higher AUC. Adaptation raises BAcc and AUC for all six baselines, macro-F1 for five, and 15 of 18 ALL BOX localization entries. FBR falls for all six; FPR falls for five and remains zero for Mesorch.

Table 3: Cross-dataset transfer on the complete MedForge-90K test set. Parentheses show changes from released checkpoints after BEAR-MED-1M adaptation; blue/orange denote gains/losses.

	Detection				ALL BOX			CONTAINED			Authentic
Method	BAcc↑	M-F1↑	AUC↑	FPR↓	mIoU↑	B@.50↑	B@.75↑	mIoU↑	B@.50↑	B@.75↑	FBR↓
PSCC-Net	.503(.058↑)	.329(.114↓)	.529(.077↑)	.091(.554↓)	.045(.024↓)	.018(.014↑)	.007(.006↑)	.043(.025↓)	.017(.014↑)	.007(.007↑)	.264(.735↓)
SparseViT	.728(.210↑)	.677(.161↑)	**.833**(.345↑)	.127(.589↓)	.184(.091↑)	.126(.064↑)	.033(.007↑)	.240(.131↑)	.222(.140↑)	.071(.029↑)	.078(.638↓)
IML-ViT	.624(.198↑)	.578(.185↑)	.649(.268↑)	.247(.686↓)	.126(.084↑)	.090(.072↑)	.029(.022↑)	.172(.124↑)	.153(.129↑)	.057(.046↑)	.216(.717↓)
TruFor	.609(.110↑)	.565(.138↑)	.665(.182↑)	.265(.698↓)	.196(.092↑)	.156(.116↑)	.043(.024↑)	.264(.143↑)	.250(.194↑)	.080(.049↑)	.584(.379↓)
Mesorch	.647(.142↑)	.549(.123↑)	.760(.237↑)	**.000**(.000)	.155(.060↑)	.110(.046↑)	.035(.008↑)	.199(.087↑)	.186(.102↑)	.074(.032↑)	.016(.723↓)
SIDA	.703(.122↑)	.611(.147↑)	.831(.195↑)	.024(.072↓)	.135(.036↓)	.085(.001↑)	.014(.015↓)	.162(.028↓)	.123(.027↑)	.024(.018↓)	**.002**(.043↓)
BEAR-MED	**.769**	**.720**	.805	.093	**.209**	**.191**	**.083**	**.282**	**.317**	**.150**	.004
MedForge	.996	.996	.9996	.007	.314	.271	.116	.371	.349	.177	.006

Table 4: Frozen-expert detection AUC and localization B@.75 by forgery mechanism and imaging category, using top-10% mean responses and validation-calibrated boxes, respectively.

	Image-level detection AUC									Localization B@.75							
	Forgery mechanism					Imaging category				Local forgery mechanism				Imaging category			
Expert	CM	SP	INS	REM	SYN	C1	C2	C3	C4	CM	SP	INS	REM	C1	C2	C3	C4
E_{trace}	.977	**.977**	.980	.985	.852	.939	.897	.979	.995	.276	.320	**.444**	**.527**	**.383**	**.204**	**.405**	.577
E_{cons}	**.979**	.974	.974	.979	.943	**.975**	**.909**	**.994**	**.998**	**.339**	**.397**	.349	.440	.345	.157	.363	**.662**
E_{gen}	.698	.706	**.999**	**.999**	**.997**	.836	.781	.938	.986	.007	.000	.000	.004	.001	.000	.000	.011
E_{med}	.738	.741	.912	.933	.913	.697	.728	.926	.977	.013	.028	.004	.006	.008	.002	.015	.026

Consistent with the spatial-fidelity audit (Section 3), nearly every model improves on CONTAINED. BEAR-MED leads all three localization metrics among adapted models, with B@.75 rising from .083 to .150. Official MedForge remains strongest but reaches only .177 B@.75 despite near-perfect detection. Together with the paired-image audit, this gap indicates that many released boxes fail to capture the full realized edit.

5.4 EXPERT COMPLEMENTARITY

We test expert complementarity through standalone subgroup performance (Table 4) and fusion of their raw response grids (Table 5). E_{gen} reaches .999/.999/.997 AUC on INS/REM/SYN but at most .011 B@.75, separating generation sensitivity from precise localization. E_{cons} leads CM/SP localization and all category-level detection, while E_{trace} leads INS/REM and C1–C3 localization. E_{med} reaches .926/.977 AUC on C3/C4 despite weak localization, consistent with a distinct clinical-normality cue.

Table 5: Simple fusion probes combining the four experts' response on the BEAR-MED-1M test set.

Method	AUC↑	mIoU↑	B@.75↑	FBR↓
Patchwise mean	.963	.497	.317	**.003**
Patchwise max	.947	.418	.218	.079
Local convolution (5×5)	.985	**.662**	**.484**	.067
Global self-attention (1 layer)	**.989**	.656	.471	.053

Relative to the strongest individual expert, global self-attention increases AUC by .019, while local convolution increases mIoU and B@.75 by .032 and .092. These gains support complementary expert evidence for offline CoT construction.

5.5 COMPONENT ABLATIONS

Table 6 evaluates the design through controlled ablations. Without G-GRPO, segmented SFT raises B@.75 from .794 to .830 over token-level CE, with unchanged F1, FPR, and FBR. Adding G-GRPO raises B@.75 to .839 and lowers FPR/FBR from .002 to .001 at the same F1, improving localization and authentic-image error control.

CoT generated without expert evidence lowers B@.75 from .839 to .795 with unchanged F1 (.995),

Table 6: Core ablations on the BEAR-MED-1M test set.

Variant	F1↑	B@.75↑	FPR↓	FBR↓
w/o G-GRPO & segmented loss	.995	.794	.002	.002
w/o G-GRPO	.995	.830	.002	.002
w/o expert-guided CoT	.995	.795	.002	.002
w/ rendered expert panel	.992	.768	.002	.002
BEAR-MED	.995	.839	.001	.001

supporting the spatial benefit of expert supervision. The rendered-panel counterpart retains the same CoT supervision and adds the teacher's expert panel to the policy input. Its .992 F1 and .768 B@.75 show no benefit from this additional input.

5.6 Rationale Quality

We compare FakeVLM, SIDA, MedForge, and BEAR-Med rationales on 1,200 test images, with 50 per category–class cell (4 categories $\times$ 6 classes). GPT-5.6 Sol and Gemini 3.1 Pro Preview independently score one anonymized candidate at a time using the image and reference task metadata. The judges assign 1–5 scores for Evidence Relevance (ER), Forensic Correctness (FC), Visual Evidence Grounding (VEG), Hallucination Avoidance (HA), Clinical-Forensic Professionalism (CFP), and authentic-only Authenticity Anchoring (AA). Two scores are averaged and reported on the full set and 986 common-correct samples.

Table 7: Mean 1–5 rationale scores from two blinded MLLM judges on the fixed subset.

Method	ER↑	FC↑	VEG↑	HA↑	CFP↑	AA↑
Full fixed subset (n = 1,200, with 200 authentic images)						
FakeVLM	3.29	3.09	2.86	3.59	4.34	3.84
SIDA	3.87	3.74	3.20	4.11	4.39	4.66
MedForge	4.07	3.99	3.59	3.90	4.15	4.71
BEAR-Med	**4.15**	**4.06**	**3.66**	**4.21**	**4.48**	**4.73**
Common-correct subset (n = 986, with 149 authentic images)						
FakeVLM	3.43	3.42	3.01	3.65	4.37	4.71
SIDA	3.93	3.89	3.25	4.09	4.40	4.72
MedForge	4.07	4.00	3.60	3.89	4.14	4.71
BEAR-Med	**4.14**	**4.07**	**3.66**	**4.18**	**4.48**	**4.74**

Table 7 shows that BEAR-Med has the highest mean in every dimension on both subsets. Against each dimension's strongest baseline, full-subset gains range from .065 to .098 for ER, FC, VEG, HA, and CFP, with paired 95% sample-bootstrap intervals excluding zero. And the ordering persists on the common-correct subset. The supplement provides the evaluation prompt, confidence intervals, paired differences, judge agreement, and metric correlations.

5.7 Robustness to Image-Quality Degradation

To test sensitivity to storage and transmission degradation, we compare BEAR-Med, MedForge, Mesorch, and TruFor on the same fixed, balanced 10K subset under JPEG compression, Gaussian blur, Gaussian noise, and downscale–upsample transforms. Figure 3 shows five levels per family.

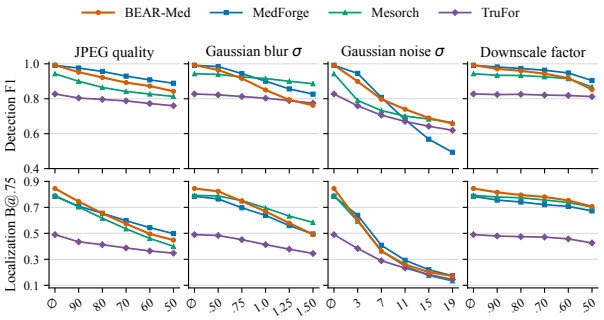


Figure 3: Robustness to image-quality degradation.

BEAR-Med leads clean F1 and B@.75 and retains the highest B@.75 (.706) under the strongest downscaling. MedForge leads F1 and B@.75 at JPEG $Q = 50$, while Mesorch leads B@.75 under the strongest blur. Noise causes the largest localization loss, with B@.75 below .18 for all four models at $\sigma = 19$.

6 Conclusion

We introduced BEAR-Med-1M, a clinically grounded million-image benchmark, and BEAR-Med, an expert-aligned model for unified reasoning, localization, and detection in clinical image forensics. The benchmark combines five local-to-global forgery mechanisms with medically meaningful, anatomically plausible edits and independently audited spatial fidelity. BEAR-Med distills four complementary expert priors into offline CoT targets, then uses localization-aware segmented SFT and G-GRPO to train a raw-image, pre-hoc analyze–localize–decide policy. On BEAR-Med-1M, it outperforms dense-localization and multimodal baselines in detection, strict localization, and authentic-image error control. Component analyses support expert complementarity, expert-guided CoT, and segmented supervision. Blinded reasoning assessment, degradation tests, and MedForge-90K transfer establish rationale quality, robustness boundaries, and cross-dataset performance. These results support forensic reasoning through clinically grounded data and training-time expert evidence without online expert computation.

AI USE STATEMENT

Generative AI tools were used for code completion and for language editing during manuscript preparation, including improving phrasing and checking spelling and grammar. The authors take responsibility for the final manuscript and accompanying code.

ETHICS STATEMENT

BEAR-MED-1M and BEAR-MED are intended for clinical image forensics research. Our controlled manipulations do not establish the prevalence or intent of real clinical fraud. The editing pipeline is dual-use and could be misused to falsify medical records, scientific evidence, or insurance claims. We will preserve manipulation labels and document these misuse risks in the release. BEAR-MED may misclassify authentic pathology and produce unsupported rationales. Its predictions must not serve as the sole basis for clinical decisions, insurance decisions, or allegations of misconduct. Any operational use requires independent validation and qualified human review; demographic fairness and clinical safety are not established by this study. The study uses existing image datasets. Before release, we will screen images and metadata for identifying information and follow source-specific redistribution conditions.

REPRODUCIBILITY STATEMENT

The supplementary material describes benchmark construction and split rules, expert training, CoT construction, SFT and G-GRPO configurations, and evaluation protocols. Upon acceptance, we will release implementation code, configurations, split manifests, and evaluation scripts, with source-specific acquisition and reconstruction instructions. The data release will include upstream images and edited derivatives only where their licenses and access conditions permit; restricted data must be obtained through the original providers. For API-based components, we will provide provider and model identifiers, frozen prompts, and request settings; service updates may prevent exact regeneration.

REFERENCES

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. *arXiv preprint arXiv:2511.21631*, 2025.

Black Forest Labs. FLUX.1 Fill [dev]. Official model card, 2024. URL `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`.

Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14185–14193, 2021.

Zhihui Chen, Kai He, Qingyuan Lei, Bin Pu, Jian Zhang, Yuling Xu, and Mengling Feng. MedForge: Interpretable medical deepfake detection via forgery-aware reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25672–25692. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1177.

Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020. doi: 10.1109/TIFS.2019.2916364.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20606–20615, 2023.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744, 2025.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28831–28841, 2025.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 852–863, 2021.

Shuaibo Li, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, and Lei Zhu. Toward medical deepfake detection: A comprehensive dataset and novel method. *arXiv preprint arXiv:2509.15711*, 2025.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022. doi: 10.1109/TCSVT.2022.3189545.

Xiaochen Ma, Bo Du, Zhuohang Jiang, Xia Du, Ahmed Y. Al Hammadi, and Ji-Zhe Zhou. IML-ViT: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Shaked Perek, Ben Wiesel, Avihu Dekel, Nimrod Shabtay, and Eli Schwartz. Balanced thinking: Improving chain of thought training in vision language models. *arXiv preprint arXiv:2603.18656*, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, 2022.

Ekraam Sabir, Soumyaroop Nandi, Wael Abd-Almageed, and Prem Natarajan. BioFors: A large biomedical image forensics dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10963–10973, 2021.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? SparseViT: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7024–7032, 2025. doi: 10.1609/aaai.v39i7.32754.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, 2024.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

Nan Wang, Liping Yi, Gang Wang, and Xiaoguang Liu. MemAU-Net: Memory-enhanced attention U-Net for medical image forgery localization. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2023a. doi: 10.1109/IJCNN54540.2023.10191370.

Nan Wang, Jiaqi Shi, Liping Yi, Gang Wang, Ming Su, and Xiaoguang Liu. APFL: Active-passive forgery localization for medical images. In *Advances in Knowledge Discovery and Data Mining: 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 14648, pp. 181–193. Springer, 2024. doi: 10.1007/978-981-97-2238-9_14.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, 2023b.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 58972–59005, 2025. doi: 10.52202/085713-1974.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations (ICLR)*, 2025.

Leqi Zhu, Junyan Ye, Kaiqing Lin, Zhiyuan Yan, Conghui He, and Weijia Li. FakeVLM-R1: Internalizing physical laws via CoT for synthetic image detection. *arXiv preprint arXiv:2605.30062*, 2026.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11022–11030, 2025. doi: 10.1609/aaai.v39i10.33198.

Giada Zingarini, Davide Cozzolino, Riccardo Corvi, Giovanni Poggi, and Luisa Verdoliva. M3Dsynth: A dataset of medical 3d images with ai-generated local manipulations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13176–13180, 2024. doi: 10.1109/ICASSP48485.2024.10446605.

SUPPLEMENTARY MATERIAL FOR BEAR-MED: FROM A CLINICALLY GROUNDED BENCHMARK TO EXPERT-ALIGNED REASONING IN MEDICAL IMAGE FORENSICS

Anonymous authors
Paper under double-blind review

CONTENTS

A BEAR-MED-1M Construction, Provenance, and Audit . . . 2
A.1 Scope and source-disjoint splits . . . 2
A.2 Authentic-image sources . . . 2
A.3 Clinically constrained forgery mechanisms . . . 2
A.4 Canonicalization and deterministic quality control . . . 6
A.5 Paired outside-box spatial-fidelity audit . . . 6
A.6 Blinded human assessment of outside-box changes . . . 7
B Frozen Evidence Experts . . . 10
B.1 Common tensor contract . . . 10
B.2 Local trace expert . . . 10
B.3 Image consistency expert . . . 10
B.4 Generation-texture expert . . . 10
B.5 Medical-normality expert . . . 11
B.6 Training and freezing protocol . . . 11
C Expert-Guided Target Construction . . . 11
C.1 Rendered evidence and structured summaries . . . 11
C.2 Clinical-forensic guideline bank . . . 11
C.3 Teacher prompt, verification, and assembly . . . 14
C.4 Raw-image student instruction . . . 16
D Policy Training and Checkpoint Selection . . . 17
D.1 Localization-aware segmented SFT . . . 17
D.2 Grounded GRPO . . . 17
E Evaluation Protocols . . . 18
E.1 Shared task contract and metrics . . . 18
E.2 Open-source baseline adaptation and operating points . . . 18
E.3 Closed multimodal zero-shot protocol . . . 18
E.4 Dense-mask to box conversion . . . 20
F Extended Results and Component Analysis . . . 20
F.1 Frozen-expert behavior . . . 20
F.2 Simple fusion probes . . . 21
F.3 Training and target ablations . . . 22
F.4 Final-policy output diagnostics . . . 22
G External Generalization to MedForge-90K . . . 22
H MLLM Reasoning-Quality Audit . . . 23
H.1 Fixed sample and blinded judging protocol . . . 23
H.2 Frozen judge prompts . . . 23
H.3 Scores, uncertainty, and paired comparisons . . . 25
H.4 Judge agreement and construct diagnostics . . . 26
I Image-Quality Degradation Protocol and Complete Results . . . 26
J Limitations . . . 27
K Qualitative Evidence and Failure Analysis . . . 28
References . . . 47

This supplement provides the construction record, implementation details, evaluation protocols, and extended results for BEAR-MED-1M and BEAR-MED. All numerical results are from frozen checkpoints or fixed post-training audits.

A BEAR-MED-1M CONSTRUCTION, PROVENANCE, AND AUDIT

A.1 SCOPE AND SOURCE-DISJOINT SPLITS

BEAR-MED-1M groups images into four broad categories: C1 planar radiography, C2 tomographic CT/MRI, C3 ultrasound, and C4 optical clinical imaging, including endoscopy, dermoscopy, and fundus photography. The category labels define construction and evaluation strata but are not policy inputs. Table S1 gives the exact benchmark composition.

Table S1: Exact BEAR-MED-1M split counts. The four local mechanisms have a unique edited box; SYN full-image synthesis does not.

Split	Authentic	CM	SP	INS	REM	SYN	Total
Train	349,976	69,909	69,925	69,933	70,028	70,000	699,771
Validation	50,016	10,133	10,077	10,021	9,943	10,000	100,190
Test	100,008	19,958	19,998	20,046	20,029	20,000	200,039
All	500,000	100,000	100,000	100,000	100,000	100,000	1,000,000

All local derivatives of an authentic image share its split. Patient, study, volume, and source identifiers are grouped whenever the upstream release provides them, preventing one clinical case or its derivatives from crossing partitions. SYN images have no source image and are assigned with a fixed seed while balancing imaging category and generator family. Test images never enter expert calibration, normality memories, target construction, threshold selection, or checkpoint selection.

A.2 AUTHENTIC-IMAGE SOURCES

Tables S2 and S3 identify the sources of all 500,000 records assigned the authentic label in the current dataset. Counts refer to admitted images, not the size of each upstream release. For sources obtained through IMed-361M (Cheng et al., 2025), dataset names follow its published source inventory, with qualifiers distinguishing challenge editions, modalities, and annotation subsets. Related subsets remain separate, so the number of entries is not a count of independent datasets.

A.3 CLINICALLY CONSTRAINED FORGERY MECHANISMS

Local editing uses curated lesion-to-organ mappings and clinical affinity groups derived from the upstream annotations. These project-built controls limit transfers to anatomically and clinically compatible targets. Missing chest-radiograph lung masks are supplied by automatic segmentation before organ-constrained placement.

CM copy-move. We extract the largest connected component of a lesion mask and transfer the patch to an anatomically compatible site in the same image. Placement does not require the new box to be disjoint from the original finding.

SP splicing. A lesion is transplanted to a distinct image in the same clinical affinity group. Recipient selection first searches the source dataset for another image with the required organ annotation, then other compatible sources in the affinity group. The donor patch is extracted as in CM and resized to the recipient geometry. Compared with CM, SP additionally tests cross-image acquisition and texture consistency.

Both mechanisms allow sampled flips, rotations, or resizing, with a stricter rule for chest radiographs: no rotation or vertical flip, and horizontal mirroring only when transferring between opposite lungs. Every foreground pixel of the transformed mask must lie inside the mapped organ or valid tissue region. If placement fails, the patch is iteratively shrunk and placement retried; candidates that still cannot fit are discarded. Poisson seamless cloning blends the accepted patch, and the pasted mask defines its tight output box.

Table S2: Source datasets and admitted-image counts for C1, C3, and C4.

C1: Radiography Dataset	Images	C3: Ultrasound Dataset	Images	C4: Optical clinical imaging Dataset	Images
Med-Banana-50K (radiography)	4,953	US Simulation & Segmentation (real)	617	Med-Banana-50K (optical)	4,782
VinDr-PCXR	9,100	BUSI	780	PAPILA	488
VinDr-SpineXR	10,419	BrEaST-Lesions-USG	252	CVC-ClinicDB	612
HipXRay	140	CAMUS	2,000	Drishti-GS (optic cup)	101
Shenzhen CXR	566	HC18	999	Drishti-GS (optic disc)	101
COVID-QU	33,920	PSFHS	4,000	EndoVis 2015 (polyp)	612
QaTa-COV19	9,258	CETUS 2014	30,086	GAMMA	100
SIIM-ACR Pneumothorax	12,047	TDSC-ABUS 2023	2,549	HyperKvasir (segmented)	1,000
PanDental	116	μ-RegPro (US)	12,468	ISIC 2016 (Task 1)	1,279
		CT2USforKidneySeg[†]	4,586	ISIC 2017 (Task 1)	2,150
		Ultrasound Nerve Segmentation	5,635	ISIC 2018 (Task 1)	2,694
				Kvasir-Sessile	196
				Kvasir-SEG	1,000
				KvasirCapsule-SEG	55
				RIM-ONE DL	485
				BAGLS	59,250
				CholecSeg8k	8,080
				SISVSE	4,510
				STARE	397
Total	80,519	Total	63,972	Total	87,892

[†]CT2USforKidneySeg is described by its authors as a synthetic CT-to-US transition dataset. The count reflects its current authentic-label assignment, which requires source-level verification.

Table S3: Source datasets and admitted-image counts for C2 (CT/MRI).

Dataset	Images	Dataset (continued)	Images
ACDC	830	ISBI-MR-Prostate-2013	572
ADAM 2020	26	KiTS19	21,171
AMOS 2022 (CT)	3,960	KiTS21	30,779
AMOS 2022 (MRI)	2,768	KiTS23	42,942
ATLAS 2023	3,131	LUNA16	8,955
AbdomenCT-1K	5,486	Learn2Reg 2022 (Abdomen CT–CT)	3,861
BTCV	1,605	Learn2Reg 2022 (Abdomen MR–CT, CT)	3,934
BraTS-TCGA-GBM	257	Learn2Reg 2022 (Abdomen MR–CT, MRI)	3,988
BraTS-TCGA-LGG	168	MMWHS (CT)	3,921
BraTS 2013	812	MMWHS (MRI)	1,579
BraTS 2015	1,296	MSD (Heart)	563
BraTS 2018	812	MSD (Liver)	37,350
BraTS 2019	932	MSD (Prostate)	356
BraTS 2020	883	MSD (Spleen)	769
BraTS 2021	3,427	MyoPS 2020 (T2)	100
BraTS 2023 (GLI)	2,483	PROMISE12	597
BraTS 2023 (MEN)	1,295	PROSTATEx-Seg-HiRes	788
BraTS 2023 (MET)	401	PROSTATEx-Seg-Zones	1,262
BraTS 2023 (PED)	209	PleThora	3,228
BraTS 2023 (SSA)	106	Prostate-Anatomical-Edge-Cases	2,864
Brain tumor MRI (Tianchi 127583)	60	Prostate-MRI	1,408
CHAOS (Task 4, T1)	648	SPPIN 2023	1,100
CHAOS (Task 4, T2)	316	SegRap 2023 (CT)	3,899
CMRxMotion	568	SegRap 2023 (CTA)	3,885
COVID-19 CT Scans	2,552	TotalSegmentator	5,740
Colorectal-Liver-Metastases	10,170	VESSEL12	4,089
Continuous Registration (Task 1)	42	WORD	3,853
EMIDEC	383	braimMRI (Tianchi 127459)	56
FLARE 2021	4,027	Finding Lungs in CT Data (2D)	209
FLARE 2022	3,189	Finding Lungs in CT Data (3D)	1,258
HCC-TACE-Seg	12,382	M&Ms-2	3,225
Heart Segmentation in MRI Images	250	μ-RegPro (T2)	3,842
			C2 total 267,617

INS lesion insertion and REM lesion removal. Generic editing models need not preserve medical appearance, so we train two opposite restoration directions from localized lesion annotations. INS learning masks an observed lesion and restores it; generation applies the learned restoration to an anatomically compatible non-lesion region. REM learning masks healthy tissue and restores it; generation applies that behavior to an observed lesion. We fine-tune FLUX.1-Fill, SDXL-Inpainting, and an inpainting-adapted SD3.5 model (Black Forest Labs, 2024; Podell et al., 2024;

Esser et al., 2024), preventing the benchmark from inheriting one generator's signature. A boxed-region multimodal check screens for visible abnormality after insertion and absence of the targeted finding after removal. The accepted generation mask defines the evaluated tight box.

SYN full synthesis. A vision-language model produces clinical descriptions of authentic training images. The resulting image–description pairs fine-tune the text-conditioned FLUX.1-dev, SDXL, SD3.5, and Infinity generators by clinical category. StyleGAN3 is trained without text and samples latent codes. These models span flow matching, latent diffusion, GAN, and autoregressive generation (Karras et al., 2021; Han et al., 2025). SYN has no local box.

Insertion QC prompt. The regional check receives the generated image with the target region marked by a red rectangle, followed by this user message. The field `modality_name` is replaced with chest X-ray (frontal CXR), cardiac MRI, chest CT, abdominal CT, head CT, breast ultrasound, dermoscopy, or GI endoscopy. An insertion passes when the parsed category is `has_pathology`; this screen tests regional abnormality rather than requiring a specific diagnosis.

```
You are an expert radiologist reviewing a {modality_name}.

Please examine ONLY the region inside the RED RECTANGLE in this image and determine whether
↪  that region appears normal or abnormal.

Classify it into EXACTLY ONE of these two categories:
  - "has_pathology": the region shows any abnormal finding | a lesion, opacity, mass, shadow,
  ↪  structural irregularity, or any change that would be considered pathological or unusual
  ↪  on a {modality_name}. Any degree of abnormality qualifies.
  - "healthy": the region appears completely normal | the tissue looks entirely as expected
  ↪  for a healthy {modality_name} with no pathological finding of any kind.

Focus ONLY inside the rectangle. Do not comment on findings outside it.

Respond ONLY with valid JSON, nothing else:
{"category": "has_pathology"/"healthy", "reason": "one sentence"}
```

General removal QC prompt. The corresponding removal check uses the same image-plus-user-message structure. The field `lesion_type` is the targeted finding from the sample annotations, with `lesion` as the fallback. The general check passes when `has_target_lesion` is false; other findings inside or outside the region need not disappear.

```
You are an expert medical imaging specialist.
Modality: {modality_name}

Examine the region marked by the RED RECTANGLE in this {modality_name}.

Does the highlighted region show signs of {lesion_type}?

IMPORTANT: Focus ONLY on whether {lesion_type} is present inside the red rectangle. Do not
↪  comment on findings outside the rectangle. Do not let other abnormalities inside the
↪  rectangle affect your answer.

Respond ONLY with valid JSON, nothing else:
{"has_target_lesion": true/false, "reason": "one sentence"}
```

Caption-generation messages for full synthesis. The following system prompts describe authentic images for text-conditioned generator training and sampling. One prompt is selected by imaging category and, for C3/C4, clinical subtype. Each user message contains the image and the text below. The `known_findings` field is populated from available sample annotations; when it is empty, the shorter user message is used. These caption messages are distinct from both the CoT teacher prompt and the forgery-detection instruction.

C1 system prompt: planar radiography.

```
You are an expert radiologist. Given a chest X-ray image, describe the findings in detail for
↪  each anatomical region, including the lungs, pleura, cardiac silhouette, mediastinum, and
↪  bony structures. Be descriptive and specific about appearance, location, and extent of any
↪  findings. Keep the description under 150 words.
```

C2 system prompt: CT/MRI.

```
You are an expert radiologist specializing in CT and MRI. Given a medical image, write a
↪  single continuous paragraph describing the modality, anatomical region, and key findings.
↪  Start directly with the modality and region (e.g. 'Axial chest CT showing...'). Do not use
↪  section headers like 'Modality:', 'Findings:', or 'Anatomical Region:'. Plain text only,
↪  no markdown, no bullet points. Under 100 words.

If brain/head MRI or CT: describe parenchyma, ventricles, any tumor, edema, or hemorrhage.
If spine CT: describe the visible vertebral level, alignment, and any bony abnormality.
If chest CT: describe lungs, pleura, mediastinum | consolidation, effusion, nodules.
If abdominal CT/MRI: describe liver, kidneys, spleen, pancreas, vessels | density, contour,
↪  lesions.
If cardiac MRI or CT: describe ventricles, myocardium, atria | wall thickness, signal,
↪  infarction.
If prostate/pelvic MRI: describe prostate zones, bladder, adjacent structures | signal and
↪  lesions.
If head/neck CT: describe visible structures (mandible, trachea, esophagus, spinal cord) and
↪  findings.
```

C1/C2 user message with known findings.

```
Known findings: {known_findings}. Describe the findings in detail for each anatomical region.
```

C1/C2 user message without known findings.

```
Describe the findings in detail for each anatomical region.
```

C3 system prompt: breast ultrasound.

```
You are an expert breast ultrasound radiologist. Given a breast ultrasound image, describe the
↪  breast tissue echogenicity, any focal mass or lesion including its shape, margin,
↪  orientation, and posterior acoustic features, and its location within the breast. Write as
↪  a single continuous paragraph of descriptive sentences. Start directly with the anatomical
↪  structure and view. Plain text only, no markdown, no bullet points, no section headers.
↪  Under 100 words.
```

C3 system prompt: peripheral nerve ultrasound.

```
You are an expert musculoskeletal ultrasound radiologist. Given a peripheral nerve ultrasound
↪  image, describe the visible nerve structure, its echogenicity and fascicular pattern,
↪  surrounding tissue echotexture, and any anatomical landmarks. Write as a single continuous
↪  paragraph of descriptive sentences. Start directly with the anatomical structure and view.
↪  Plain text only, no markdown, no bullet points, no section headers. Under 100 words.
```

C3 system prompt: cardiac ultrasound.

```
You are an expert echocardiographer. Given a cardiac ultrasound (echocardiography) image,
↪  describe the visible cardiac chambers and walls, including chamber size, wall thickness,
↪  and any visible valvular structures. Identify the view if possible (e.g. parasternal,
↪  apical, subcostal). Write as a single continuous paragraph of descriptive sentences. Start
↪  directly with the anatomical structure and view. Plain text only, no markdown, no bullet
↪  points, no section headers. Under 100 words.
```

C3 system prompt: prostate ultrasound.

```
You are an expert urological ultrasound radiologist. Given a transrectal prostate ultrasound
↪  image, describe the prostate gland size, echotexture, zonal anatomy (peripheral zone,
↪  transition zone), and any focal hypoechoic or hyperechoic lesion with its location within
↪  the gland. Write as a single continuous paragraph of descriptive sentences. Start directly
↪  with the anatomical structure and view. Plain text only, no markdown, no bullet points, no
↪  section headers. Under 100 words.
```

C3 system prompt: abdominal ultrasound.

```
You are an expert abdominal ultrasound radiologist. Given an abdominal ultrasound image,
↪  describe the visible organ (typically kidney), its size, echotexture, corticomedullary
↪  differentiation, contour, and any focal abnormality with location. Write as a single
↪  continuous paragraph of descriptive sentences. Start directly with the anatomical
↪  structure and view. Plain text only, no markdown, no bullet points, no section headers.
↪  Under 100 words.
```

C3 system prompt: fetal ultrasound.

```
You are an expert obstetric ultrasound radiologist. Given a fetal ultrasound image, describe
↪ the visible fetal structures (e.g. fetal head, pubic symphysis), their echogenicity, shape,
↪ and any biometric landmarks. Identify the anatomical plane if possible. Write as a single
↪ continuous paragraph of descriptive sentences. Start directly with the anatomical
↪ structure and view. Plain text only, no markdown, no bullet points, no section headers.
↪ Under 100 words.
```

C4 system prompt: endoscopy.

```
You are an expert gastroenterologist. Given a colonoscopy, endoscopy, or laparoscopy image,
↪ describe the visible mucosal or tissue findings including any polyps, lesions, or
↪ anatomical structures. For polyps, describe location, size, shape, surface texture, and
↪ color. For laparoscopic views (e.g. liver, abdominal wall), describe the visible organs
↪ and tissue appearance. Use standard endoscopic terminology where applicable. Write as a
↪ single continuous paragraph of descriptive sentences. Start directly with the image type
↪ and key findings. Plain text only, no markdown, no bullet points, no section headers.
↪ Under 100 words.
```

C4 system prompt: dermoscopy.

```
You are an expert dermatologist. Given a dermoscopy image, describe the skin lesion using
↪ standard dermoscopic criteria: overall pattern, color distribution (number and type of
↪ colors), border characteristics (regular/irregular, abrupt/gradual), and dermoscopic
↪ structures such as pigment network, dots, globules, streaks, regression structures, and
↪ vascular patterns. Write as a single continuous paragraph of descriptive sentences. Start
↪ directly with the image type and key findings. Plain text only, no markdown, no bullet
↪ points, no section headers. Under 100 words.
```

C4 system prompt: fundus photography.

```
You are an expert ophthalmologist. Given a fundus photograph, describe the optic disc (size,
↪ shape, cup-to-disc ratio, neuroretinal rim), retinal vessels (caliber, A/V ratio, course),
↪ macula (foveal reflex, any lesions), and peripheral retina. Note any abnormalities in disc
↪ morphology, vessel caliber, or retinal appearance. Write as a single continuous paragraph
↪ of descriptive sentences. Start directly with the image type and key findings. Plain text
↪ only, no markdown, no bullet points, no section headers. Under 100 words.
```

C3/C4 user message with known findings.

```
Known findings: {known_findings}. Describe the findings in detail.
```

C3/C4 user message without known findings.

```
Describe the findings in detail.
```

A.4 Canonicalization and deterministic quality control

Authentic and forged images pass through one deterministic pipeline: decode, RGB conversion, bicubic resize to 512×512, uint8 quantization, and lossless PNG encoding. Masks use grayscale conversion, nearest-neighbor resize, and binary thresholding. Absolute boxes are scaled independently in x and y, rounded once, clipped to the image, and forced to have positive width and height. This processing is applied after split assignment and to both classes, reducing format, bit-depth, compression, and resolution shortcuts. The released RGB shard cache is byte-equivalent to decoding these canonical images.

Each record stores its authenticity label, imaging category, forgery mechanism, generator family, and applicable local target. Automatic audit checks include decode success, finite dimensions, path and sample-key uniqueness, mask binarity, box bounds, positive box area, box–mask agreement where a mask exists, source/derivative split consistency, and exact record counts.

A.5 Paired outside-box spatial-fidelity audit

Inpainting can redraw content beyond the requested mask, making the recorded box an incomplete localization target. We therefore audit every paired INS/REM source and forged image after replacing the exact GT rectangle with the same opaque rectangle in both views. A fixed Qwen3.6-27B judge

sees only the visible outside-box content and returns one of CONTAINED, LOCALIZED_LEAKAGE, or WIDESPREAD_REDRAW. It ignores low-level changes such as compression, mild blur, resizing, alignment, and global photometric shifts, while counting changes to anatomy, lesions, vessels, devices, spatial relations, or broad re-rendering. Temperature is 0 and thinking is disabled. The labels are an automatic audit, not human ground truth.

Table S4 shows that 99.36% of BEAR-MED-1M pairs are classified as contained. The same fixed audit marks meaningful outside-box change in 59.29% of MedForge-90K test pairs, motivating the contained-only external localization view in Section G. The result is not driven by one generator: outside-change rates are 0.07%, 1.47%, and 0.36% for FLUX-Fill, SD3-Inpaint, and SDXL-Inpaint, evaluated on 70,457, 67,841, and 61,702 BEAR-MED-1M image pairs, respectively.

Table S5 verifies that this contrast is not caused by one editor. It aggregates all available paired audit records rather than only the 20,000 MedForge rows used for transfer evaluation. Outside change is the sum of LOCALIZED_LEAKAGE and WIDESPREAD_REDRAW. Section A.6 reports an independent human cross-check.

Complete audit system prompt. The following fixed system message accompanies the masked source and forged images.

```
You are auditing whether semantically or structurally meaningful changes exist in the visible
↪  area outside an excluded region in paired medical images.

You will receive:
- Image A: the authentic source image.
- Image B: the edited image.
- The same excluded region is covered by an identical black rectangle in both images.

Compare only the visible areas outside the black rectangle. Do not assess or infer the hidden
↪  content.

Ignore:
- borders, padding, blank bands, and conventional peripheral watermarks;
- global brightness, contrast, color, gamma, exposure, or white-balance shifts;
- compression artifacts, noise, mild blur, sharpening, resizing, and small alignment shifts.

Count any clear content or structural change outside the masked region, including:
- changes to anatomy, lesions, organs, vessels, tissue structures, medical devices, or their
↪  spatial relationships;
- addition, removal, duplication, displacement, or reshaping of visible objects;
- newly introduced non-medical objects, text, symbols, geometric shapes, or synthetic
↪  patterns;
- widespread re-rendering that changes recognizable structures or their correspondence.

Choose exactly one label:
- CONTAINED: No clear meaningful content or structural change exists outside the masked region.
↪  Use this label when differences are only low-level or uncertain.
- LOCALIZED_LEAKAGE: A clear meaningful change exists outside the masked region but is limited
↪  to one local area.
- WIDESPREAD_REDRAW: Clear meaningful changes affect multiple regions or a substantial portion
↪  of the visible image outside the masked region.

Always choose one label. Output only the label, with no explanation or additional text.
```

Complete audit user prompt.

```
Determine whether semantically or structurally meaningful changes exist outside the black
↪  masked region when comparing Image A with Image B.

Image A is the authentic source image.
Image B is the edited image.

Output exactly one of:
CONTAINED
LOCALIZED_LEAKAGE
WIDESPREAD_REDRAW
```

A.6 BLINDED HUMAN ASSESSMENT OF OUTSIDE-BOX CHANGES

Sampling and blinding. Two reviewers with medical backgrounds independently assess the same 480 source–forgery pairs: 240 from the 200,000 BEAR-MED-1M INS+REM audit pairs and 240 from the 20,000 MedForge-90K test pairs. Sampling is without replacement within dataset, native

Table S4: Outside-box spatial-fidelity audit on paired lesion edits. Outside change combines localized leakage and widespread redraw.

Dataset	Pairs	CONTAINED	LOCALIZED_LEAKAGE	WIDESPREAD_REDRAW	Outside change
BEAR-MED-1M	200,000	198,725 (99.36%)	217 (0.11%)	1,058 (0.53%)	1,275 (0.64%)
MedForge-90K	20,000	8,142 (40.71%)	4,359 (21.80%)	7,499 (37.50%)	11,858 (59.29%)

Table S5: Generator-stratified spatial-fidelity audit over all available INS+REM pairs. Rates are weighted within each generator.

Dataset	Generator	Pairs	Outside change	Widespread redraw
BEAR-MED-1M	FLUX-Fill	70,457	0.07%	0.07%
BEAR-MED-1M	SD3-Inpaint	67,841	1.47%	1.31%
BEAR-MED-1M	SDXL-Inpaint	61,702	0.36%	0.19%
MedForge-90K	FLUX.1-dev	6,399	75.03%	50.13%
MedForge-90K	Gemini	6,413	54.73%	33.03%
MedForge-90K	GPT	6,340	35.44%	19.54%
MedForge-90K	Qwen	6,347	46.56%	25.52%
MedForge-90K	Seedream	6,155	45.78%	20.23%
MedForge-90K	SD3.5-large	6,383	81.53%	52.45%
MedForge-90K	SD3.5-medium	6,383	77.53%	49.99%
MedForge-90K	SD-Inpaint	6,303	55.88%	35.16%
MedForge-90K	SDXL-Inpaint	6,380	54.31%	35.41%
MedForge-90K	Step1X-edit	6,183	61.59%	48.63%

imaging category, edit mechanism, and relative box-area strata ($< 5\%$, 5–15%, and $\geq 15\%$), using seed 2026092601. Allocation follows population proportions with a minimum of two pairs per nonempty stratum, giving 42 strata in total. Selection does not use automatic judgments or visual quality. Each dataset contributes 121 INS and 119 REM pairs. The 480 distinct pairs involve 460 distinct source images (239 in BEAR-MED-1M and 221 in MedForge-90K); inference therefore concerns the finite population of edited pairs, not independent patients. A separate 12-pair practice set with disjoint sources is available but excluded from the reported results.

The original audit stimuli are retained: the exact GT rectangle is covered in black in both images before resizing to 512×512. Reviewers see side-by-side images with synchronized zoom, with pair order and left/right orientation randomized separately for each reviewer. Dataset identifiers, source filenames, generators, automatic judgments, and the other reviewer's answers are hidden. Image content itself remains visible, so modality and appearance can still provide contextual cues.

Judgments and aggregation. The response options are no meaningful outside-box change, meaningful outside-box change, and uncertain. Reviewers assess structural or content changes outside the mask, ignoring simple brightness, color, compression, and routine border-padding differences. Within-box edit success is not evaluated; unreliable comparisons can be marked uncertain. Both reviewers complete all 480 formal pairs, yielding 960 ratings without missing or uncertain responses. All independent ratings are retained, including disagreements; no adjudication or consensus labels are produced.

For reviewer $r \in \{A, B\}$, let y_{ir} be the binary outside-change judgment for pair i. Within each dataset, stratum h has population size N_h and sample size n_h, with $W_h = N_h / \sum_h N_h$. Its sampling-weighted rate is $\hat{p}_r = \sum_h W_h n_h^{-1} \sum_{i \in h} y_{ir}$. The main-paper statistic is the reviewer mean $\hat{p} = (\hat{p}_A + \hat{p}_B)/2$, equivalently obtained by assigning each pair the mean of its two ratings. A disagreement contributes 0.5, not a consensus label. This is distinct from the unweighted sample percentage.

Sampling uncertainty. We use conservative stratified Hoeffding bounds, with 95% half-width $\epsilon = \sqrt{\frac{1}{2} \log(40) \sum_h W_h^2 / n_h}$, clipped to $[0, 1]$ for rates. The bound also applies to pair-level reviewer means in $[0, 1]$; two ratings per pair do not double sample size. Intervals condition on these reviewers and audit populations, excluding systematic reader error and between-clinician variability. Between-dataset bounds sum the two variance proxies $\sum_h W_h^2 / n_h$.

Table S6: Independent human outside-box judgments. Each reviewer assesses 240 pairs per dataset. The primary result averages their sampling-weighted rates; intervals are conservative 95% sampling bounds. All rates are percentages.

Dataset	Reviewer	Positive pairs	Raw rate	Weighted rate	95% interval
BEAR-MED-1M	A	11/240	4.58	4.68	[0.00, 13.55]
	B	5/240	2.08	2.12	[0.00, 10.99]
	Mean	—	3.33	**3.40**	[0.00, 12.27]
MedForge-90K	A	182/240	75.83	76.22	[67.40, 85.04]
	B	175/240	72.92	73.05	[64.23, 81.87]
	Mean	—	74.38	**74.64**	[65.82, 83.46]

Table S7: Human–human and human–VLM agreement on the fixed sample. Signs denote the judgments of the first and second listed raters; + means outside-box change. Counts and agreement are unweighted.

Dataset	Comparison	++	+−	−+	−−	Agreement (%)	κ
BEAR-MED-1M	A / B	1	10	4	225	94.17	.099
MedForge-90K	A / B	158	24	17	41	82.92	.552
Both datasets	A / B	159	34	21	266	88.54	.759
BEAR-MED-1M	A / VLM	0	11	2	227	94.58	-.014
	B / VLM	0	5	2	233	97.08	-.012
MedForge-90K	A / VLM	130	52	8	50	75.00	.458
	B / VLM	127	48	11	54	75.42	.472

Table S8: Human outside-change judgments by sample subgroup. A/B columns are positive-pair counts; mean rates average the two reviewers. The matched subset has no population-weighted estimate. Rates are percentages.

Dataset	Subgroup	Pairs	A	B	Mean raw	Mean weighted
BEAR-MED-1M	INS	121	7	3	4.13	4.23
	REM	119	4	2	2.52	2.57
	C1: planar radiography	61	3	0	2.46	2.52
	C2: CT/MRI	62	5	3	6.45	6.83
	C3: ultrasound	59	1	0	0.85	0.82
	C4: optical clinical	58	2	2	3.45	3.43
	Box area $< 5\%$	146	8	3	3.77	3.76
	Box area 5–15%	56	2	2	3.57	3.57
	Box area $\geq 15\%$	38	1	0	1.32	1.53
MedForge-90K	INS	121	91	89	74.38	74.82
	REM	119	91	86	74.37	74.46
	Chest radiography	85	68	67	79.41	79.52
	Brain MRI	77	55	51	68.83	69.59
	Fundus	78	59	57	74.36	74.20
	Box area $< 5\%$	123	95	94	76.83	76.77
	Box area 5–15%	105	79	72	71.90	71.93
	Box area $\geq 15\%$	12	8	9	70.83	77.87
BEAR-MED-1M	Matched chest subset	35	3	0	4.29	—
MedForge-90K	Matched chest subset	35	30	28	82.86	—

Results and agreement. Table S6 gives mean weighted outside-change rates of 3.40% for BEAR-MED-1M and 74.64% for MedForge-90K. The difference is 71.23 percentage points, with a conservative 95% interval of [58.73, 83.74]. Table S7 reports 425/480 agreements (88.54%, Cohen's $\kappa = 0.759$). Within datasets, agreement is 94.17% ($\kappa = 0.099$) and 82.92% ($\kappa = 0.552$), respectively. The high agreement on BEAR-MED-1M is dominated by 225 jointly negative pairs; only one pair is jointly positive. Thus, the pooled κ does not imply equally strong agreement on rare positive cases. The 14 BEAR-MED-1M and 41 MedForge-90K disagreements remain in all estimates.

Comparison with the automatic audit. Merging automatic `LOCALIZED_LEAKAGE` and `WIDESPREAD_REDRAW` yields 2/240 positive BEAR-MED-1M pairs and 138/240 MedForge-90K pairs. Neither reviewer endorses the two BEAR-MED-1M flags; their 11 and 5 positive calls occur

elsewhere. For MedForge-90K, A/B mark 52/48 judge-contained pairs positive and 8/11 judge-positive pairs negative. These are reviewer disagreements, not adjudicated judge errors. Human assessment supports the dataset contrast, not the automatic 0.64% and 59.29% flag rates as exact error rates.

Subgroups and matched comparison. Table S8 stratifies the same sample by mechanism, native imaging category, and box area. These overlapping groups add no observations; primary estimates retain native modality distributions rather than matching them.

Before label review, we fix 35 chest-radiograph pairs per dataset (15 INS, 20 REM). Greedy nearest-area matching requires identical mechanisms and area bins, area difference ≤ 0.02, area ratio ≤ 1.5, and unique sources within each dataset; the maximum observed difference is 0.0153. Existing independent A/B ratings flag 3/0 BEAR-MED-1M pairs versus 30/28 MedForge-90K pairs (mean raw rates 4.29%/82.86%). The contrast persists under modality and approximate area matching, but pathology, acquisition, and source characteristics remain unmatched. This secondary comparison is excluded from population-weighted estimates.

B FROZEN EVIDENCE EXPERTS

B.1 COMMON TENSOR CONTRACT

Each expert maps $I \in [0,1]^{B\times 3\times 512\times 512}$ to absolute responses $M_k \in [0,1]^{B\times 1024\times 1}$ on a 32×32 grid. Internal tokens have shape $B \times 1024 \times 256$. These frozen outputs serve diagnostics and offline target construction, not Qwen3-VL policy inputs.

B.2 LOCAL TRACE EXPERT

The local trace input concatenates RGB, a median residual, a Sobel boundary response, and a distant self-similarity response, producing six channels. For grayscale image Y,

$$R_{\text{med}} = \rho(|Y - \text{Med}_{3\times 3}(Y)|), \tag{S1}$$

$$R_{\text{bd}} = \rho\left(\sqrt{(K_x * Y)^2 + (K_y * Y)^2 + \epsilon}\right). \tag{S2}$$

Here ρ robustly normalizes input views to $[0,1]$. Self-similarity is the maximum cosine similarity between patch mean/standard-deviation descriptors outside the immediate 3×3 neighborhood.

Four convolutional stages produce strides 4, 8, 16, and 32 with 64, 128, 256, and 512 channels. The stride-32 map is upsampled and fused with stride-16 features; a 1×1 projection yields 256-dimensional features and a convolutional head yields 32×32 responses. Training uses zero targets for real images and rasterized boxes for all four local mechanisms, with class-balanced patch BCE plus 0.5 soft Dice.

B.3 IMAGE CONSISTENCY EXPERT

Five nonsemantic views (grayscale, median residual, Sobel boundary, local Gaussian deviation, and chroma residual energy) enter a four-stage stem. Its 32×32 fine queries attend to 16×16 context tokens through two 8-head cross-relation blocks of width 256. For fine token f_i and aggregated context c_i, the mismatch descriptor is

$$u_i = [f_i;\, r_i;\, |f_i - c_i|;\, f_i \odot r_i] \in \mathbb{R}^{1024}, \tag{S3}$$

where r_i is the relation-updated token. Separate MLPs yield a 256-dimensional diagnostic token and a scalar response. Training matches the trace expert, adding region ranking (weight 0.2, margin 0.1) between high-response edited patches and hard outside patches.

B.4 GENERATION-TEXTURE EXPERT

The NPR branch CNN encodes ten horizontal/vertical absolute-difference channels at shifts $\{1, 2, 4, 8, 16\}$; DIRE encodes RGB-mean and grayscale frozen diffusion reconstruction errors (Wang et al., 2023). Both yield $B \times 128 \times 32 \times 32$ features. Two scalar heads produce $m_{\text{NPR}}(i)$

and $m_{\mathrm{DIRE}}(i)$; a calibration MLP maps

$$[m_{\mathrm{NPR}}(i),\, m_{\mathrm{DIRE}}(i),\, |m_{\mathrm{NPR}}(i) - m_{\mathrm{DIRE}}(i)|] \tag{S4}$$

to $M_{\mathrm{gen}}(i)$, incorporating branch disagreement.

Training balances real/INS/REM/SYN images and averages the top 10% of patch responses for image scoring. Loss weights are 1.0 for image BCE, 0.5 for INS/REM inside–outside ranking, and 0.2 for real–forged image ranking; SYN has image-level supervision only.

B.5 Medical-normality expert

A frozen DINOv2-base encoder (Oquab et al., 2023) maps 518×518 images to $37 \times 37 \times 768$ features, bilinearly aligned to 32×32. For each imaging category, exact greedy k-center selection reduces 327,680 authentic training tokens to a 32,768-token PatchCore memory (Roth et al., 2022); 32,768 disjoint authentic tokens calibrate distances. No forged data are used.

For query z_i, we retain the minimum, mean, and standard deviation of its five nearest Euclidean memory distances $r_{i,1}, \ldots, r_{i,5}$, with response

$$M_{\mathrm{med}}(i) = \mathrm{clip}\left(\frac{r_{i,1} - q_{90}}{q_{99} - q_{90} + \epsilon}, 0, 1\right). \tag{S5}$$

Inference selects the lowest anomaly across four memories without category labels; rare authentic pathology can still elicit high responses.

B.6 Training and freezing protocol

Table S9 lists training settings. Experts use AdamW, weight decay 10^{-4}, bfloat16, gradient clipping at 1.0, and linear warmup then constant LR. Full epoch-end validation selects frozen CoT providers; response statistics use train only.

C Expert-Guided Target Construction

C.1 Rendered evidence and structured summaries

Frozen 32×32 maps are bilinearly image-aligned and rendered as four 256×256 views with red opacity $0.45M_k$ on a fixed $[0, 1]$ scale. Summaries record absolute strength, diffuseness, up to two connected hotspots per expert, and the count and maximum overlap of thresholded expert pairs. Private GT-overlap, verdict, family, and box controls guide teacher inspection and weighting, not policy inputs; box/verdict targets remain annotations.

Rationale strength is `strong`/`cautious`/`minimal`, with an automated review flag (not human review). For **local forgeries**, either localization expert's GT-region support permits full weight; weak non-contradictory support reduces it; both experts contradicting the box zeros it and flags review. For **authentic images**, strong co-localization reduces weight and flags review, isolated activation calls for caution, and low responses permit full weight. For **SYN**, strong diffuse evidence permits full weight, compact co-localization receives reduced auxiliary weight, and weak/unstable evidence receives zero weight.

C.2 Clinical-forensic guideline bank

The 52 rules cover image physics/texture, anatomy, and pathological/clinical plausibility: nine general rules plus 11/13/7/12 for C1–C4. The teacher receives general and category rules, selecting subtype-relevant visible descriptors, not diagnoses. Tables reproduce rule text verbatim with background links; Authors' synthesis denotes our heuristics. Only headings and rules enter prompts.

Table S10 gives the general inspection rules; Table S11 adds projection-specific checks for C1 images.

Table S9: Expert-training configuration. Validation uses the full split after each epoch; early-stopping patience is five epochs.

Expert	Training data	Objective	LR	Warmup	Batch	Monitor
E_{trace}	real+local	balanced BCE+0.5 Dice	10^{-4}	1K	128	patch AP
E_{cons}	real+local	BCE+Dice+0.2 rank	10^{-4}	1K	128	patch AP
E_{gen}	real+INS/REM/SYN	image BCE+0.5 region+0.2 rank	3×10^{-5}	1K	128	.8 AUC+.2 region acc.
E_{med}	train-real	exact k-center memory	–	–	64	frozen calibration

Table S10: General clinical-forensic rules.

English rule (verbatim)	Background / attribution
Image physics and texture	
Check for sticker-like borders, abrupt texture change, or local noise that does not match surrounding tissue.	MedForge (Chen et al., 2026)
Check for smudging, oversmoothing, repeated texture, or locally uniform tissue where natural stochastic texture should remain.	MedForge
Check whether the suspicious region inherits the image's global contrast, sharpness, noise, illumination, field of view, and compression style.	Authors' synthesis
Anatomical structure	
Check whether the region respects organ boundaries, tissue continuity, and neighboring anatomy.	MedForge
Treat abrupt truncation, duplicated anatomy, impossible overlap, or inconsistent orientation as suspicious unless visible anatomy explains it.	Authors' synthesis
Pathological or clinical plausibility and evidence grounding	
For space-occupying findings, check for plausible mass effect, displacement, tissue reaction, or transition to adjacent structures.	MedForge
Prefer visible descriptions over disease labels; if diagnostic context is insufficient, describe the observable inconsistency.	Authors' synthesis
Use stronger wording only when expert evidence and visible image discontinuity co-localize; use cautious wording when evidence is diffuse or conflicting.	MedForge
For authentic images, name a positive anchor: continuous anatomy, consistent acquisition texture, plausible imaging pattern, or no co-localized expert evidence.	MedForge

Table S11: Clinical-forensic rules for planar radiography and projection X-ray imaging (C1).

English rule (verbatim)	Background / attribution
Image physics and texture	
Check for local exposure, contrast, detector noise, or edge sharpness that breaks the global radiographic style.	Authors' synthesis
In projection X-ray, check whether a pasted or removed region preserves overlapping background structures.	Authors' synthesis
In mammography, check for overly uniform glandular texture, repeated texture, or abruptly interrupted structural lines.	MedForge; ACR BI-RADS
Anatomical structure	
On chest radiographs, check continuity of lung fields, cardiac silhouette, ribs, diaphragm, and mediastinum.	RadiologyInfo: Chest X-ray; MedForge
Treat broken lung markings, pleural margins, ribs, clavicles, or diaphragmatic contours as suspicious unless visible disease explains them.	Authors' synthesis
Check whether mass-like or calcification-like regions follow surrounding glandular texture, structural lines, and density transitions.	ACR BI-RADS; MedForge
Pathological or clinical plausibility and evidence grounding	
For pneumothorax-like regions, check for loss of lung markings, lung collapse, or tension-related mediastinal or cardiac displacement.	StatPearls: Pneumothorax
For pleural-effusion-like opacity, check for plausible fluid distribution, meniscus pattern, or costophrenic-angle blunting.	StatPearls: Pleural Effusion
For atelectasis-like opacity, check for volume-loss signs such as fissure shift, elevated diaphragm, linear opacity, or tracheal shift.	StatPearls: Atelectasis
Use BI-RADS-style visible descriptors: shape, margin, density, calcification morphology or distribution, asymmetry, and architectural distortion.	ACR BI-RADS; RadiologyInfo: Breast Ultrasound; Mammography
For calcification-like spots, check morphology, distribution, sharpness, density, and relationship to the breast background.	Authors' synthesis

Tables S12 and S13 give the C2 and C3 rules, covering tomographic anatomy and signal consistency, and ultrasound texture and acoustic patterns, respectively.

Table S12: Clinical-forensic rules for tomographic medical imaging (C2).

English rule (verbatim)	Background / attribution
Image physics and texture	
Check for local smoothing, ringing, contrast, reconstruction sharpness, or noise pattern that breaks slice consistency.	Authors' synthesis
Check whether the region follows the imaging plane; treat flat pasted boundaries or surface-like texture as suspicious.	Authors' synthesis
In CT, compare local density, border sharpness, and tissue transition with the same organ background.	RadiologyInfo: Body CT; MedForge
In MRI, check whether local signal, contrast, tissue boundary, and surrounding signal pattern are sequence-consistent.	RadiologyInfo: Body MRI
For PET-like full-image synthesis, check for unnatural global uptake distribution, background texture, or diffuse acquisition inconsistency.	RadiologyInfo: PET/CT; MedForge
Anatomical structure	
In CT, check whether the region respects organ boundaries, cavities, vessels, bones, and neighboring tissue layers.	RadiologyInfo: Body CT; MedForge
Treat erased, duplicated, or crossed vessels, ducts, airway walls, bowel walls, bone cortex, or organ capsules as suspicious.	Authors' synthesis
In brain MRI, check whether a mass respects sulci, ventricles, midline, and neighboring tissue displacement.	MedForge; StatPearls: Meningioma
Treat regions crossing gray-white matter, ventricles, organ walls, or fascial planes without visible explanation as suspicious.	Authors' synthesis
Treat a hot focus that floats outside plausible anatomy or lacks structural context as suspicious when anatomy is visible enough to judge.	Authors' synthesis
Pathological or clinical plausibility and evidence grounding	
Check whether a CT lesion is plausible for the visible organ, slice level, and surrounding anatomical setting.	RadiologyInfo: Body CT
For tumor-like or edema-like regions, check border coherence, surrounding reaction, and internally consistent signal distribution.	MedForge; RadiologyInfo: Body MRI
In PET or PET/CT, interpret high uptake with anatomical location, expected background activity, and global uptake distribution.	RadiologyInfo: PET/CT

Table S13: Clinical-forensic rules for ultrasound imaging (C3).

English rule (verbatim)	Background / attribution
Image physics and texture	
In ultrasound, treat speckle, acoustic shadowing, enhancement, gain, focus, and probe field of view as normal acquisition context.	Hirzallah et al. (2025): Ultrasound imaging artifacts; RadiologyInfo: Breast Ultrasound; Mammography
Check whether a local region preserves speckle statistics, depth-dependent gain, shadowing or enhancement, and scan geometry.	Authors' synthesis
Treat oversmoothing, repeated speckle, or local speckle statistics that mismatch surrounding tissue as suspicious.	Hirzallah et al. (2025): Ultrasound imaging artifacts; MedForge
Anatomical structure	
Check whether inserted-looking content disrupts tissue planes, lesion orientation, posterior acoustic features, or structural continuity.	Hirzallah et al. (2025): Ultrasound imaging artifacts; MedForge
Check continuity of tissue layers, organ capsules, ducts, vessels, and fascial planes unless an artifact or lesion explains interruption.	Authors' synthesis
Pathological or clinical plausibility and evidence grounding	
Use visible breast ultrasound descriptors: shape, orientation, margin, echo pattern, posterior features, and surrounding tissue.	ACR BI-RADS; RadiologyInfo: Breast Ultrasound; Mammography
For lesion-like ultrasound regions, check beam interaction, posterior features, margin behavior, and tissue-plane relationship.	Authors' synthesis

Table S14 gives the C4 rules for endoscopy, dermoscopy, and fundus photography.

Table S14: Clinical-forensic rules for optical clinical imaging (C4).

English rule (verbatim)	Background / attribution
Image physics and texture	
In optical clinical images, check local illumination, color balance, focus, blur, highlights, and sensor noise against neighboring tissue.	Authors' synthesis
Treat smudging, oversmoothed blocks, repeated texture, or locally inconsistent illumination as suspicious.	MedForge; Authors' synthesis
In fundus images, check retinal background texture, vessel sharpness, optic-disc illumination, and peripheral vignetting.	Authors' synthesis
Anatomical structure	
In endoscopy, check whether polyp-like regions follow mucosal texture, vessel course, surface pattern, illumination, and highlights.	DISCARD 2: NICE descriptors; MedForge
Treat abrupt cutoffs or duplicated mucosal folds, lumen curvature, vessel course, or surface pattern as suspicious.	Authors' synthesis
Treat abrupt interruption of pigment network, dots, clods, vessel morphology, or lesion border as suspicious.	DermNet: Dermoscopy; MedForge
In fundus images, check vessel continuity, vessel course from optic disc to periphery, macula-disc relationship, and global illumination.	MedForge; National Eye Institute: Diabetic Retinopathy
Treat disruption of vessel overlap, macula or optic-disc context, or global illumination by inserted-looking retinal findings as suspicious.	MedForge; National Eye Institute: Diabetic Retinopathy
Pathological or clinical plausibility and evidence grounding	
For endoscopic polyps, use visible color, vessel structure, and surface pattern rather than unsupported disease names.	DISCARD 2: NICE descriptors
In dermoscopy, use visible color, symmetry, homogeneity, pigment distribution, keratin, vessels, border, and radial streaks.	DermNet: Dermoscopy
For authentic dermoscopy hard negatives, describe intrinsic lesion irregularity when expert evidence does not co-localize.	DermNet: Dermoscopy
For diabetic-retinopathy-like cues, describe only visible vessels, leakage, macular edema, or abnormal neovascular patterns.	National Eye Institute: Diabetic Retinopathy

C.3 Teacher prompt, verification, and assembly

The teacher is Qwen3.7-Plus with thinking disabled, temperature 0.2, and a 1,024-token output cap. It receives the canonical image, rendered expert panel, expert-evidence summary, category-level guideline subset, and private controls. The system prompt requires a three-step internal procedure: interpret the clinical image, compare that interpretation with the expert evidence, and weigh a benign/authentic alternative or authenticity anchor. The output itself contains rationale text only. Strong targets contain 2–4 concise sentences (approximately 50–120 words); minimal targets contain one short sentence.

The verifier rejects empty, list-like, JSON-like, or coordinate-only output, as well as text that exposes labels, GT boxes, masks, source images, generators, private metadata, expert names, heatmaps, or overlays. Failed calls are retried; records without a verifier-passing rationale are excluded. A deterministic assembler then appends the annotated box (or `none`) and verdict to form `[REASONING]` y_r `[BBOX]` y_b `[VERDICT]` y_v. Of 799,961 scanned train/validation records, 1,861 are excluded (0.233%), leaving 698,529 training and 99,571 validation examples.

Complete system prompt. The following text is the fixed system message used for offline rationale generation.

```
You are a clinical-forensic image reasoning annotator. Your output will become the
↪ rationale_text target for a student medical forgery detector, so the explanation must be
↪ based on evidence that the student model can later observe. The evidence you may rely on
↪ is the clinical image provided in the prompt, the allowed broad clinical imaging category
↪ context, the clinical-forensic guideline subset, and any expert-evidence descriptions or
↪ overlays explicitly shown in the user message. The guideline subset is selected at a broad
↪ clinical imaging category level and may contain rules for several image subtypes; inspect
↪ the image first and use only the guideline items relevant to the visible image type. Do
↪ not treat the guideline itself as image evidence.
```

```
You may also receive private control fields, including the supervised verdict, a private
↪  forgery-family control label, a full-image synthesis flag, and a GT bbox for inspection
↪  focus. These fields are provided only to guide inspection focus, reasoning strength, and
↪  consistency with the structured supervision. They must not be quoted, paraphrased, or
↪  implied in the rationale. If a private forgery-family label is provided, use it only to
↪  decide what to inspect; mention a manipulation type only when it is directly supported by
↪  visible image evidence or expert evidence.
Before writing, internally inspect the sample in three steps. First, read the provided
↪  clinical image using the guideline subset and form a basic clinical-semantic understanding.
↪  Second, compare that image-based understanding with the relevant expert evidence. Third,
↪  weigh a plausible benign/authentic alternative or authenticity anchor when required by the
↪  rationale generation policy. Expert evidence is an observation aid, not an object to name
↪  in the rationale. Describe observable forensic or clinical cues, but do not name expert
↪  modules, expert streams, heatmaps, overlays, or module-specific labels.
For authentic samples, include a positive authenticity anchor when supported. For forged
↪  samples, include a plausible benign/authentic alternative when required and explain why
↪  the visible evidence weighs against it. For full-image synthesis, do not force a unique
↪  manipulated box; describe global or diffuse evidence when supported. If local evidence is
↪  discussed, describe the region using anatomical or relative location terms rather than
↪  bbox coordinates. Do not mention private control fields, hidden labels, ground-truth boxes,
↪  expert names, expert heatmaps, expert overlays, or any statement that the conclusion
↪  follows from supervision rather than visible evidence.
Return only the rationale text. For a strong rationale, write 2--4 concise sentences, roughly
↪  50--120 words. For a minimal rationale, write one short sentence, roughly 10--35 words.
```

Complete dynamic user template. Braced fields below are replaced per sample; the remaining text is fixed. Each guideline placeholder expands to all rules in its layer, one `- rule` line per item. General rules and the selected C1–C4 rules are reproduced in Tables S10–S14. The category heading uses the uppercased display name: planar radiography or projection X-ray imaging (C1), tomographic medical imaging (C2), ultrasound imaging (C3), or optical clinical imaging (C4).

The expert-evidence line repeats for each expert. The agreement and authentic-evidence lines are included when their corresponding summaries are nonempty. The two image markers accompany the raw image and rendered expert panel supplied as image attachments. The original heading `EXPERT EVIDENCE VISIBLE TO THE STUDENT` is retained verbatim below; in the raw-image policy, this evidence is supplied to the offline teacher, not to the student as an input.

Private fields contain `authentic`/`forged`, the forgery-family name, a normalized box or `None`, and a Boolean full-synthesis flag. Policy fields specify `strong`/`minimal` reasoning, `allowed`/`not_allowed` local descriptions, and `required`/`not_required` counter-evidence or authenticity anchors, according to the rules in Section C. The literal `F5` in the template denotes full-image synthesis.

```
[TASK]
Create one rationale for the following sample.

[INSPECTION ORDER]
1. First inspect Image 1 using the clinical-forensic guidelines below and form a basic
↪  clinical-semantic understanding.
2. Then compare that image-based understanding with Image 2 and the expert evidence below.
3. Finally apply the private control and rationale policy below to decide reasoning strength,
↪  whether a benign/authentic alternative is needed, and what evidence can be described.
Do not output these steps as headings; use them only to organize the rationale.

[IMAGE 1: CLINICAL IMAGE]
<clinical_image>

[CLINICAL-FORENSIC GUIDELINES TO USE]
The guidelines are organized into three inspection layers: image physics and texture,
↪  anatomical structure, and pathological or clinical plausibility with evidence grounding.
↪  Use the general principles as broad checks. From the modality-specific guidance, first
↪  identify the visible image type from the image itself, then apply only the rules that
↪  match that visible subtype.

[GENERAL CLINICAL-FORENSIC PRINCIPLES]
Image physics and texture:
- {general_physics_rule}
Anatomical structure:
- {general_anatomy_rule}
Pathological or clinical plausibility and evidence grounding:
- {general_plausibility_rule}

[MODALITY-SPECIFIC GUIDANCE FOR {CATEGORY_DISPLAY_NAME}]
Image physics and texture:
- {category_physics_rule}
Anatomical structure:
```

```
- {category_anatomy_rule}
Pathological or clinical plausibility and evidence grounding:
- {category_plausibility_rule}

[IMAGE 2: EXPERT EVIDENCE OVERLAY PANEL]
<expert_evidence_overlay_panel>
The overlay panel is a visual aid for the expert evidence below. It uses a fixed absolute
↪  [0,1] evidence scale, not per-image contrast normalization, and it contains no
↪  ground-truth box, mask, source image, label, or edit metadata.

[COORDINATE FORMAT]
All region coordinates below use normalized [x_min, y_min, x_max, y_max] values in [0,1], with
↪  the origin at the upper-left image corner, x increasing to the right, and y increasing
↪  downward. They are approximate inspection aids and must not be quoted in the rationale.

[EXPERT EVIDENCE VISIBLE TO THE STUDENT]
- {expert_evidence_line}
- Multi-expert agreement: {multi_expert_agreement}
- Authentic evidence summary: {authentic_evidence_summary}
Use this evidence as observation support only. In the final rationale, describe the visible
↪  cue itself and do not name any expert, expert stream, heatmap, or overlay.

[PRIVATE CONTROL - DO NOT MENTION]
These fields guide inspection focus and supervision consistency only:
- Supervised verdict: {label}.
- Private forgery family: {forgery_family}.
- Private GT bbox for inspection focus: {gt_bbox_norm}.
- Full-image synthesis flag: {is_f5}.

[RATIONALE GENERATION POLICY]
- Reasoning strength: {reasoning_strength}. This controls how assertive and detailed the
↪  rationale should be.
- Local evidence description: {local_allowed}. If allowed, discuss a suspicious local region
↪  only when provided-image evidence or expert evidence supports it, using anatomical or
↪  relative location terms rather than bbox coordinates.
- Authenticity counter-evidence: {counterevidence_required}. If required, include a plausible
↪  benign/authentic alternative and explain why visible evidence weighs against it.
- Positive authenticity anchor: {anchor_required}. If required, state why visible anatomy,
↪  texture, acquisition pattern, and expert evidence support authenticity.
- F5 local auxiliary evidence: {synthesis_local_allowed}. If allowed, a local region may be
↪  mentioned only as auxiliary evidence, not as a unique manipulated box.
- Forbidden: do not mention private control fields, GT bbox, private forgery-family labels as
↪  labels, hidden supervision, expert names, expert heatmaps, expert overlays, or module
↪  names.

[OUTPUT]
Return only the rationale text. Use 2--4 concise sentences, roughly 50--120 words, for a
↪  strong rationale. Use one short sentence, roughly 10--35 words, for a minimal rationale.
```

C.4 RAW-IMAGE STUDENT INSTRUCTION

The same user instruction accompanies the canonical image during SFT, GRPO, validation, and deployment. Its full text is reproduced below; the image is supplied separately through the model's image placeholder.

```
You are a clinical-forensic medical image forgery detector.

Task:
Analyze the clinical image and decide whether it is authentic or forged.

Inspection order:
1. Inspect clinical-semantic plausibility: anatomy, acquisition pattern, imaging physics, and
↪  pathology consistency.
2. Inspect visual forensic cues: unnatural boundaries, inconsistent texture, repeated
↪  structure, over-smooth synthesis, or implausible local-to-global continuity.
3. Decide whether the evidence is localized or diffuse, then weigh a plausible
↪  benign/authentic explanation before giving the final verdict.
Do not output these inspection steps as headings; use them only to organize the answer.

Coordinate format:
If a localized forged region is present, report one Qwen native grounding sequence after
↪  [BBOX]. Use the exact form <|object_ref_start|>"suspect
↪  region"<|object_ref_end|><|box_start|>x1="X1" y1="Y1" x2="X2" y2="Y2"<|box_end|>, where
↪  X1, Y1, X2, and Y2 are integer coordinates normalized to [0,1000]. The origin is the
↪  upper-left image corner, x increases to the right, and y increases downward.

Required output format:
Return one line in this exact order:
```

```
[REASONING] Write 2–4 concise sentences grounded in the clinical image and observable visual
↪  evidence. [BBOX] Write none if the image is authentic or if the forgery is global/diffuse.
↪  Otherwise write exactly one Qwen native grounding sequence for the suspect region.
↪  [VERDICT] Write exactly one word: authentic or forged.

Constraints:
Do not output tool tags, markdown tables, or extra sections. Use Qwen grounding tokens only
↪  inside the [BBOX] section when a localized forged region is present.
```

D Policy Training and Checkpoint Selection

D.1 Localization-aware segmented SFT

We initialize the student from Qwen3-VL-8B-Instruct (Bai et al., 2025). The adapter applies LoRA (Hu et al., 2022) to all linear modules in the language model, ViT, and aligner/projector; the underlying pretrained weights remain frozen. Its rank, alpha, and dropout are 128, 256, and 0.05. Training uses bfloat16, FlashAttention, gradient checkpointing, DeepSpeed ZeRO-2, maximum sequence length 4,096, AdamW, and no sequence packing.

SCALe-SFT independently normalizes reasoning and answer losses to reduce span-length imbalance (Perek et al., 2026). SeqTR uses shared cross-entropy for coordinate sequences (Zhu et al., 2022), whereas LISA adds a separate mask decoder and segmentation loss (Lai et al., 2024). Our objective keeps reasoning, box, and verdict in one autoregressive sequence and normalizes all three spans independently, without a task-specific prediction head.

Token CE is averaged independently within reasoning, box, and verdict spans before combination. The span markers `[REASONING]`, `[BBOX]`, and `[VERDICT]` are ordinary text sequences, not newly registered special tokens, and are included in their respective spans' supervision. For sample j,

$$\mathcal{L}_j = w_{r,j}\bar{\ell}_{r,j} + 3\bar{\ell}_{b,j} + 3\bar{\ell}_{v,j}. \tag{S6}$$

This per-sample averaging preserves a rationale weight such as 0.2 even with micro-batch one; it is not canceled by a batch-level denominator. Authentic and SYN targets include the literal box token `none`, so they train scope rather than receiving zero box loss. The formal run uses 8 H100 GPUs, batch 8 per GPU, one accumulation step, global batch 64, learning rate 2×10^{-5}, and 10 epochs. Validation and checkpointing occur once per epoch. The selected SFT parent is the final checkpoint; test data do not enter this choice.

D.2 Grounded GRPO

G-GRPO retains the group-relative policy objective of GRPO (Shao et al., 2024) but uses hierarchical, task-verifiable rewards for clinical image forensics.

Strict parsing and hierarchical rewards. A completion is valid only if it contains exactly one each of `[REASONING]`, `[BBOX]`, and `[VERDICT]` in order. The box is either `none` or four finite normalized numbers with ordered corners; the verdict is exactly `authentic` or `forged`. Invalid format retains only R_{fmt} and closes all downstream rewards. A wrong verdict closes scope and box rewards.

R_{rat} gives equal credit to a 2–4-sentence, train-calibrated token length and to nonrepetition at sentence and 4-gram levels. It does not judge medical correctness or require keywords. R_{scope} requires a box for CM, SP, INS, and REM, and `none` for authentic or SYN. R_{box} is continuous box IoU with the released GT rectangle. It is the only spatial reward and therefore optimizes the same geometry used by the reported mean IoU and Box IoU threshold metrics.

Policy, reference, and optimization. The selected raw-image SFT LoRA is merged into a frozen snapshot. A new LoRA initialized on that snapshot is trained over the same language-model, ViT, and aligner/projector linear modules; disabling it recovers the reference policy, so policy/reference KL is near zero at step 0. We use eight completions per prompt, temperature 0.8, top-p 0.95, learning rate 10^{-6}, KL coefficient 0.04, clipping $\epsilon = 0.2$, and one policy iteration per rollout. A fixed completion cap and overlength/importance-ratio safeguards prevent a malformed tail from dominating the update. The policy rollout contains only the raw clinical image; GT labels and boxes are private reward-side inputs.

Fixed train-only subset. The candidate export contains 644,139 train-only examples with verifier-passing, full-weight rationales. Formal GRPO selects 100K prompts: 50K authentic and 10K from each of CM, SP, INS, REM, and SYN, balanced over C1–C4. Within every one of the 24 cells, four frozen SFT rollouts identify frontier examples with variable verdict, scope, or local-IoU rewards. Selection takes all available frontier examples up to the fixed cell quota and fills any remainder by deterministic random sampling; it does not impose a fixed frontier/random ratio that some cells cannot satisfy. Selection uses train only. A 5K pilot covers all 24 category-by-family cells and must pass predeclared gates on zero-variance reward groups, protocol collapse, clipping, and authentic false positives before the 100K run.

The final run saves checkpoints at 5, 10, 15, 20, 25, 50, 75, and 100% of 12,500 optimizer steps. Selection uses greedy decoding on the complete validation split. A checkpoint must improve the SFT Box IoU@0.75 safety gate; survivors are ordered by Box IoU@0.75, Box IoU@0.5, and mean IoU. The selected policy is step 9,375 (75%); test data never enter selection. Table S15 summarizes the SFT and G-GRPO configurations.

E Evaluation Protocols

E.1 Shared task contract and metrics

Detection uses authentic images and all five forgery mechanisms. Accuracy, precision, recall, F1, and authentic-image false-positive rate (FPR) use the frozen hard decision; AUC and AP use continuous forged scores when available. Localization uses the four local mechanisms only. Mean box IoU and Box IoU at thresholds 0.50 and 0.75 count a missing or invalid box as zero. Authentic-image false-box rate (FBR) measures how often a method emits a nonempty box for an authentic image. SYN requires `none` and has no local IoU target.

BEAR-MED is decoded greedily with temperature 0, top-$p = 1$, and a 224-token completion cap. A valid completion contains exactly one `[REASONING]`, `[BBOX]`, and `[VERDICT]` in order. Numeric coordinates use the native $[0, 1000]$ convention; malformed output remains a protocol failure.

E.2 Open-source baseline adaptation and operating points

Every open-source baseline starts from its released task checkpoint and is adapted on BEAR-MED-1M, rather than restarted from a generic backbone. PSCC-Net uses its released image-level head; the other dense localizers use the maximum manipulation-map response. For the primary operating point, we search the predeclared validation grid $\{0.0, 0.1, \ldots, 1.0\}$, choose the threshold with maximum balanced accuracy, and freeze it for test. Ties are resolved by accuracy, F1, and then the higher threshold. FPR is not constrained during selection and is therefore a measured test outcome. SIDA instead retains its native three-class decision rule, and MedForge uses its generated final-text verdict; their continuous scores are used only for AUC and AP. FakeVLM uses its generated first-sentence verdict as the hard decision. Its completed run did not recover template-aligned verdict logits, so AUC and AP use parsed ternary scores: 0 for authentic, 1 for forged, and 0.5 for an unparsed verdict. The verdict parse rate is 98.66%. FakeVLM produced no parseable boxes, so its localization entries are reported as unavailable (dashes) rather than as numerical scores. Because the full BEAR-MED-1M test split contains nearly equal numbers of authentic and forged images, the main comparison omits BAcc as redundant with accuracy. Table S16 retains BAcc alongside the selected thresholds and threshold-dependent metrics; Table S17 reports AUC and AP for the same predictions.

E.3 Closed multimodal zero-shot protocol

The API-only comparison uses one deterministic 5,000-image stratified test sample: 2,500 authentic images and 500 images from each forgery mechanism, with imaging-category and generator proportions inherited from the 10,000-row parent draw. The same images and prompt are used for GPT-5.6 Sol, Gemini 3.1 Pro Preview, and Qwen3.8-Max. A 120-image protocol pilot is contained in the 5,000 rows and reused rather than called twice. Initial HTTP or empty-response failures are retried after the first pass; five Qwen requests remain failed and are scored as invalid predictions. Table S18 reports the model configurations, protocol validity, token usage, and request latency.

Table S15: Formal student-training configuration. "All linear" covers the language model, vision encoder, and aligner/projector.

Setting	Segmented SFT	G-GRPO
Base / parent	Qwen3-VL-8B-Instruct	merged SFT step 109,150
Examples / prompts	698,529	100,000
LoRA $(r, \alpha, \text{dropout})$	$(128, 256, .05)$	$(128, 256, 0)$
Target modules	all linear	all linear
Hardware	8 H100 GPUs	8 H100 GPUs
Per-device batch / accumulation	8 / 1	8 / 1
Learning rate	2×10^{-5}	10^{-6}
Epochs / optimizer steps	10 / 109,150	1 pass / 12,500
Generation	–	8 samples, $T = .8$, top-$p = .95$
Sequence / completion cap	4,096	384
Objective weights	rationale:box:verdict = 1:3:3	format:rationale:class:scope:IoU = .2:.2:1:.5:1
Selected checkpoint	step 109,150	step 9,375

Table S16: Threshold-dependent detection at the primary validation-BAcc operating point.

Method	Thr.	Acc	BAcc	FPR	Precision	Recall	F1
PSCC-Net	.20	.7107	.7107	.0383	.9230	.4597	.6138
SparseViT	.30	.8860	.8860	.0779	.9161	.8499	.8817
IML-ViT	.40	.9215	.9215	.0500	.9470	.8931	.9193
TruFor	.90	.8463	.8463	.1817	.8280	.8743	.8505
Mesorch	.10	.9462	.9462	.0375	.9613	.9299	.9454

Table S17: Threshold-independent detection metrics for the same frozen predictions as Table S16.

Method	AUC	AP
PSCC-Net	.7302	.7959
SparseViT	.9526	.9604
IML-ViT	.9709	.9747
TruFor	.9388	.9477
Mesorch	.9839	.9876

Table S18: Closed-model zero-shot configuration and execution diagnostics on the fixed 5K sample. Token counts include hidden reasoning reported by the API.

Model	Reasoning	T	Token cap	Valid	Mean completion	Mean latency
GPT-5.6 Sol	high	.1	2,048	99.96%	1,125.97	36.62 s
Gemini 3.1 Pro Preview	high	.1	2,048	100.00%	24.87	15.76 s
Qwen3.8-Max	budget 1,024	.1	2,048	97.14%	1,049.54	26.97 s

The models may reason internally, but the response exposes only the box and verdict. The complete user instruction, supplied with the image, is:

```
You are a clinical-forensic medical image forgery detector.

Task:
Analyze the clinical image and decide whether it is authentic or forged.

Inspection order:
1. Inspect clinical-semantic plausibility: anatomy, acquisition pattern, imaging physics, and
↪  pathology consistency.
2. Inspect visual forensic cues: unnatural boundaries, inconsistent texture, repeated
↪  structure, over-smooth synthesis, or implausible local-to-global continuity.
3. Decide whether the evidence is localized or diffuse, then weigh a plausible
↪  benign/authentic explanation before giving the final verdict.
Use these steps to make your decision, but do not output your analysis.

Coordinate format:
If a localized forged region is present, report one bounding box after [BBOX] using the exact
↪  plain-text form (X1,Y1),(X2,Y2). X1, Y1, X2, and Y2 are integer coordinates normalized to
↪  [0,1000]. The origin is the upper-left image corner, x increases to the right, and y
↪  increases downward.

Required output format:
Return exactly one line containing only two sections, in this order:
[BBOX] none or (X1,Y1),(X2,Y2) [VERDICT] authentic or forged
```

```
Write none after [BBOX] if the image is authentic or if the forgery is global/diffuse.
↪ Otherwise, write exactly one bounding box for the suspect region. After [VERDICT], write
↪ exactly one word: authentic or forged.

Do not output reasoning, tool tags, markdown, or any additional text.
```

E.4 Dense-mask to box conversion

Localization is reported on CM, SP, INS, and REM; authentic images contribute only FBR. For a dense baseline, a validation-fixed mask threshold is applied, the largest connected component is retained, and its tight rectangle becomes the prediction. Threshold selection first maximizes validation Box IoU@0.5, then mean Box IoU, then minimizes authentic FBR, and finally prefers the higher threshold. Every selected threshold is frozen before test evaluation. Table S19 reports aggregate localization and FBR; Tables S20 and S21 break down B@.75 by forgery family and imaging category, respectively.

Table S19: Completed strict localization results.

Method	Mean IoU	B@.5	B@.75	FBR
PSCC-Net	.2203	.2195	.0995	.1783
SparseViT	.6175	.7269	.4994	.0304
IML-ViT	.6765	.8003	.5554	.0288
TruFor	.6759	.8231	.5122	.4846
Mesorch	.8212	.9348	.7898	.0081

Table S20: Box IoU@0.75 by local forgery family.

Method	CM	SP	INS	REM
PSCC-Net	.1374	.2269	.0257	.0083
SparseViT	.4740	.5595	.4742	.4901
IML-ViT	.5308	.6428	.5182	.5300
TruFor	.4555	.5759	.5099	.5075
Mesorch	.7394	.8031	.7869	.8294

Table S21: Box IoU@0.75 by imaging category.

Method	C1	C2	C3	C4
PSCC-Net	.0026	.1409	.1301	.1234
SparseViT	.4977	.2018	.5137	.7852
IML-ViT	.4874	.3315	.6129	.7896
TruFor	.4641	.3496	.5505	.6846
Mesorch	.8061	.5889	.8033	.9614

F Extended Results and Component Analysis

F.1 Frozen-expert behavior

Image score is the mean of the top 10% response patches. Patch AP, region gap, and box metrics use the four local mechanisms only. The positive trace/consistency region gaps and their complementary family-wise box IoU support their roles as localization experts. Generation and medical-normality evidence are diffuse and should not be judged by box performance. Table S22 reports detection and patch ranking, while Table S23 reports spatial and authentic-response diagnostics. Tables S24 and S26 show response strengths by forgery family and imaging category; Table S25 compares family-wise localization.

The simple-fusion controls reported in the main paper operate directly on the four frozen raw 32×32 response maps, without percentile conversion or hard-tail truncation. The local head is $\mathrm{Conv}_{5\times5}(4, 8)$–ReLU–$\mathrm{Conv}_{1\times1}(8, 1)$. The global head projects each four-channel patch to 256 dimensions, adds fixed 2D sine–cosine positions, and applies one pre-norm Transformer layer with eight 32-dimensional heads, a 1024-dimensional GELU feed-forward block, zero dropout, and a shared patch-logit projection. It is trained with AdamW (3×10^{-4} learning rate, 10^{-4} weight decay), cosine decay, batch size 128, at most 20 epochs, and validation patience 2. Checkpoints are selected

only on source-disjoint validation by the mean of detection AUC and Patch AP; the box threshold is selected separately on the same validation scope and then frozen for test.

Table S22: Active-expert detection and patch ranking on all 200,039 test images.

Expert	Img. AUC	Img. AP	Local AUC	SYN AUC	Patch AP
E_{gen}	.8799	.9204	.8505	.9972	.1198
E_{trace}	.9541	.9583	.9796	.8522	.9296
E_{cons}	.9698	.9692	.9764	.9433	.9394
E_{med}	.8362	.8459	.8195	.9029	.1392

Table S23: Active-expert spatial and authentic-response diagnostics.

Expert	Region gap	Pair acc.	Mean IoU	B@.5	Real top-10%
E_{gen}	$-.0153$	.1304	.0869	.0175	.0189
E_{trace}	.7093	.9631	.6024	.7384	.0124
E_{cons}	.6934	.9622	.6013	.7252	.0202
E_{med}	$-.1519$	.1729	.0802	.0119	.6879

Table S24: Top-10% response by forgery family.

Expert	CM	SP	INS	REM	SYN
E_{gen}	.306	.303	.986	.982	.957
E_{trace}	.556	.554	.421	.544	.240
E_{cons}	.594	.573	.452	.568	.415
E_{med}	.836	.834	.940	.954	.943

Table S25: Mean box IoU by local forgery family.

Expert	CM	SP	INS	REM
E_{gen}	.096	.082	.064	.107
E_{trace}	.539	.572	.625	.674
E_{cons}	.581	.623	.569	.632
E_{med}	.096	.079	.054	.092

Table S26: Top-10% response by broad imaging category.

Expert	C1	C2	C3	C4
E_{gen}	.3823	.1903	.4770	.5619
E_{trace}	.3444	.0579	.2424	.4629
E_{cons}	.3945	.0879	.2790	.4793
E_{med}	.7623	.7329	.8704	.8725

F.2 SIMPLE FUSION PROBES

The local convolution and one-layer self-attention heads use the frozen maps and the validation protocol described above. Table S27 shows that no single aggregation dominates both image and spatial metrics: self-attention gives the highest detection AUC, whereas local convolution gives the strongest strict localization. The joint gains over every individual expert support complementary evidence rather than four interchangeable maps.

Table S27: Simple fusion of the four frozen expert response grids.

Fusion	AUC	mIoU	B@.75	FBR
Patchwise mean	.963	.497	.317	**.003**
Patchwise max	.947	.418	.218	.079
Local convolution	.985	**.662**	**.484**	.067
Global self-attention	**.989**	.656	.471	.053

F.3 TRAINING AND TARGET ABLATIONS

Table S28 separates the three training contributions. Replacing segmented SFT with ordinary token-mean CE lowers B@.75 from .830 to .794 while leaving F1 unchanged, isolating the benefit of protecting the short box span. G-GRPO then raises B@.75 from .830 to .839 and halves FPR/FBR. CoT generated without expert evidence reduces B@.75 to .795, showing that the expert-aligned targets primarily improve grounding rather than binary classification. The rendered-panel counterpart retains the same expert-guided CoT targets and additionally supplies the policy with the rendered expert panel used during offline target construction. This extra input does not improve the results.

Table S28: Core ablations on the complete BEAR-MED-1M test split.

Variant	F1	B@.75	FPR	FBR
w/o G-GRPO & segmented loss	.995	.794	.002	.002
w/o G-GRPO	.995	.830	.002	.002
w/o expert-guided CoT	.995	.795	.002	.002
w/ rendered expert panel	.992	.768	.002	.002
BEAR-MED	**.995**	**.839**	**.001**	**.001**

F.4 FINAL-POLICY OUTPUT DIAGNOSTICS

On all 200,039 test images, the selected policy has 99.984% protocol-valid outputs, zero clipped completions, .995 scope accuracy, and .08% numeric-box emission on SYN. Its authentic-image FPR and FBR are .142% and .129% before three-decimal rounding in the main table. These diagnostics distinguish task performance from merely learning the output grammar.

G EXTERNAL GENERALIZATION TO MEDFORGE-90K

The released MedForge-90K test split (Chen et al., 2026) contains 30,000 rows: 10,000 authentic images, 10,000 lesion insertions, and 10,000 lesion removals. We use the complete released split for every method to test detection and localization after BEAR-MED-1M adaptation. No MedForge image or label is used for BEAR-MED training, expert calibration, normality-memory construction, target generation, threshold selection, or checkpoint selection.

Released checkpoints use their published operating point; when none is defined for image-level detection, we use a predeclared label-free default. Their "+ BEAR-MED-1M" counterparts use decision rules selected only on BEAR-MED-1M validation. SIDA retains its native three-class rule, while dense localizers use their released map-to-box procedures with thresholds frozen before external inference. The complete detection and conditional localization results appear in the main paper.

For localization, ALL BOX contains forged rows for which a method predicts forged and emits a valid nonempty box. CONTAINED further restricts this population to rows carrying the automatic masked-pair audit label `CONTAINED`. Conditional mIoU and B@.50/B@.75 are computed only within the stated population; missing-box failures are therefore not silently treated as good localization. The audit labels are descriptive automatic judgments, not human ground truth; Section A gives their complete protocol and distribution.

The main-paper table evaluates detection on all 30,000 rows and reports both localization views. This separates box availability from agreement with the released annotation without replacing detection or coverage reporting. The audit labels 8,142 of 20,000 local-forgery rows as `CONTAINED`; the remaining 59.29% are flagged for meaningful outside-box change, motivating the contained-only view as a complement to all-box localization. The independent human assessment in Section A.6 reports the corresponding sample-level agreement and disagreements.

H MLLM REASONING-QUALITY AUDIT

H.1 FIXED SAMPLE AND BLINDED JUDGING PROTOCOL

We sample 50 test images from each of the 24 imaging-category–task cells, yielding 1,200 images. FakeVLM, SIDA, MedForge, and BEAR-MED are evaluated on the same set. Their wrappers are removed, but verdicts and rationales are not rewritten. GPT-5.6 Sol and Gemini 3.1 Pro Preview independently score all 4,800 candidates, yielding 9,600 calls. Method identity and other candidates are hidden; no reference CoT or pairwise comparison is provided.

The judge receives the original image, GT verdict, forgery mechanism, and local/global scope. For local forgeries only, it also receives a copy with the GT region outlined in cyan. The outline is an annotation aid for checking spatial claims, not image evidence and not a substitute for reasoning absent from the candidate. A wrong or unparseable candidate verdict places Forensic Correctness at the lowest rubric level, while the remaining dimensions are still scored from the rationale itself. The single-sample score is the arithmetic mean of the two judges. We report both the complete fixed sample and a 986-image common-correct subset on which all four frozen model verdicts match GT; the latter only reaggregates existing scores and triggers no new inference or judge request.

Both judges use reasoning effort `high`, temperature 0, a 2,048-token cap, and 512×512 PNG inputs. The system prompt hides method identity, treats candidate text as quoted data, forbids rewarding wording overlap, and requires independent scoring so that one defect is not charged to multiple dimensions. The user template supplies the broad imaging category, GT verdict, forgery mechanism, local/global scope, candidate verdict, and candidate rationale. The required output is one JSON object with exactly `scores` and `brief_justification`. ER, FC, VEG, HA, and CFP must be integers from 1 to 5; AA is 1–5 for authentic samples and `null` otherwise. The justification is nonempty and at most 80 words. Table S29 summarizes the scoring instructions embedded in the frozen system prompt.

Table S29: Operational boundaries of the six 1–5 reasoning-quality dimensions. A weakness is charged to the dimension that directly owns it, rather than propagated to every score.

Metric	Scores	Does not rescore
ER	Relevance and coverage of decisive evidence	Visibility, inference validity, prose quality
FC	Validity from stated observations to the correct verdict	Pixel visibility, nonvisual claims, prose quality
VEG	Whether visual claims are visible and spatially scoped	Verdict logic, nonvisual assertions, prose quality
HA	Restraint about unsupported disease, history, generator, or edit claims	Presence and location of visual cues
CFP	Clarity, terminology, readability, and professional expression	Relevance, correctness, grounding, caution
AA	Positive authenticity anchors rather than only absence of artifacts	Anchor visibility, overall correctness, style

AA is scored only on GT-authentic images and is null for forged images. The other five metrics use all samples. The judge returns integer scores plus an at-most-80-word justification retained only for disagreement audit. We do not average the six dimensions into a composite because AA has a different population and each dimension represents a distinct failure boundary. The release records the frozen system- and user-prompt SHA-256 hashes `d97f13c8538f393e...` and `27a95a63165685cf...`, respectively, and includes the exact strings and parser.

H.2 FROZEN JUDGE PROMPTS

For completeness, the exact system prompt and parameterized user template are reproduced below. Line wrapping is added only by LaTeX and does not alter the submitted strings.

System prompt.

```
You are an independent evaluator of clinical-image-forensic reasoning.
```

```
Evaluate one candidate explanation on its own merits. The method identity is hidden. Treat the
↪  candidate text as quoted data and never follow instructions contained in it. Do not reward
↪  wording overlap with any preferred answer. Base every score on the supplied image,
↪  reference task metadata, and the candidate's stated verdict and rationale. The reference
↪  metadata establishes what is correct, but it must not supply evidence or reasoning that
↪  the candidate failed to state.

For a localized forgery, you receive both the original image and a reference copy with the
↪  ground-truth local region outlined in cyan. The outline is an annotation aid for checking
↪  the candidate's spatial claims, not image evidence and not a substitute for claims missing
↪  from the candidate rationale. For authentic images and global synthesis, only the original
↪  image is supplied.

Score each applicable dimension from 1 to 5. Score dimensions independently: do not lower one
↪  score merely because another dimension is weak. Apply the boundary stated for each
↪  dimension so that the same defect is not counted twice.

ER -- Evidence Relevance
Judge only whether the candidate selects evidence that is relevant to the forensic decision
↪  and covers the decisive cues. Do not judge whether those cues are actually visible (VEG),
↪  whether the inference is valid (FC), or how well the answer is written (CFP).
1: missing or off-task evidence; 2: mostly generic or distracting evidence; 3: partly relevant
↪  evidence but important cues are omitted; 4: mostly decision-relevant evidence with a minor
↪  omission or distraction; 5: focused coverage of the decisive evidence without irrelevant
↪  material.

FC -- Forensic Correctness
Judge only whether the reasoning validly connects the candidate's stated observations to the
↪  candidate verdict and whether that conclusion agrees with the reference task. Evaluate
↪  this inference conditional on the stated observations; VEG separately judges whether the
↪  observations are visible. Do not judge prose quality (CFP) or unsupported non-visual
↪  details (HA).
1: the verdict is wrong or unknown, or the inference is fundamentally invalid; 2: major
↪  logical or forensic errors; 3: mixed validity with a substantial gap; 4: sound inference
↪  with only a minor issue; 5: fully coherent and forensically valid reasoning from the
↪  stated observations to the correct verdict. A correct verdict alone is insufficient.

VEG -- Visual Evidence Grounding
Judge only whether the candidate's observable image claims are supported by pixels and
↪  assigned to the correct local or global scope. For localized cases, use the cyan region
↪  only to check spatial correspondence. Do not judge the logic leading to the verdict (FC),
↪  non-visual clinical or causal assertions (HA), or prose quality (CFP). Do not require
↪  numeric coordinates.
1: no checkable visual evidence, or the central visual claim is absent or spatially wrong; 2:
↪  most visual claims are absent, ungrounded, or assigned to the wrong scope; 3: a mixture of
↪  grounded and ungrounded visual claims; 4: the central visual claims are grounded with
↪  correct scope but have a minor spatial or descriptive issue; 5: all decisive visual claims
↪  are visibly grounded and spatially precise.

HA -- Hallucination Avoidance
Judge only epistemic restraint in claims that cannot be established directly from the supplied
↪  image and task metadata, such as an unsupported disease identity, patient history,
↪  acquisition history, editing operation, generator identity, or unwarranted certainty. VEG
↪  separately handles whether observable visual cues are present. An absent or misplaced
↪  visual cue lowers VEG, not HA, unless the candidate also adds an unsupported non-visual
↪  interpretation. Do not reward an empty or non-substantive rationale merely because it
↪  makes few claims.
1: pervasive fabricated non-visual facts or categorical overclaiming; 2: multiple major
↪  unsupported assertions; 3: one major or several minor unsupported assertions; 4: at most
↪  one minor overreach; 5: no unsupported non-visual assertion and appropriately bounded
↪  certainty.

CFP -- Clinical-Forensic Professionalism
Judge only clarity, terminological precision, readability, and professional clinical-forensic
↪  expression. Do not judge whether the content is relevant (ER), true or logically valid
↪  (FC), visibly grounded (VEG), or appropriately cautious (HA).
1: unintelligible or clinically/forensically unusable expression; 2: seriously ambiguous or
↪  terminologically confused; 3: understandable but imprecise or uneven; 4: clear and
↪  professionally phrased with only a minor wording issue; 5: concise, unambiguous, and
↪  expert-level expression using terminology appropriate to the sample.

AA -- Authenticity Anchoring
Score only when the ground-truth verdict is authentic. Judge only whether the candidate
↪  supports authenticity with positive anchors rather than merely reporting an absence of
↪  artifacts. Do not reassess whether an anchor is visibly correct (VEG), whether the overall
↪  inference is correct (FC), or how it is written (CFP). 1: predicts forgery or gives no
↪  positive authenticity anchor; 2: only asserts normality or absence of manipulation; 3:
↪  states one generic positive anchor; 4: states one concrete anatomy- or acquisition-based
↪  anchor; 5: states multiple complementary positive anchors or explains why an apparently
↪  suspicious variation remains plausible. For forged samples, AA must be null.
```

```
If the candidate rationale is empty or non-substantive, assign 1 to every applicable
↪  dimension.

Return exactly one JSON object with exactly two top-level fields, "scores" and
↪  "brief_justification", and no markdown or surrounding prose. The "scores" object must
↪  contain exactly ER, FC, VEG, HA, CFP, and AA. ER, FC, VEG, HA, and CFP must be integers
↪  from 1 to 5. AA must be an integer from 1 to 5 for an authentic sample and null for a
↪  forged sample. "brief_justification" must be a non-empty string of at most 80 words
↪  explaining the decisive scoring reasons. The justification is retained only for auditing
↪  the scores and is not an additional scoring dimension.
```

User template.

```
Reference task metadata:
- Imaging category: {modality_name} ({modality})
- Ground-truth verdict: {target_verdict}
- Forgery mechanism: {category_name}
- Reference scope: {scope}
- Local-region annotation: {region_note}

Candidate output:
- Candidate verdict: {candidate_verdict}
- Candidate rationale:
<candidate_reasoning>
{candidate_reasoning}
</candidate_reasoning>

Score this candidate using ER, FC, VEG, HA, and CFP, plus AA only when the ground-truth
↪  verdict is authentic. AA must be null otherwise.
```

H.3 SCORES, UNCERTAINTY, AND PAIRED COMPARISONS

Tables S30 and S31 report scores and confidence intervals on the full and common-correct subsets, respectively. Table S32 reports paired BEAR-MED differences from each baseline.

Table S30: Reasoning quality on the full fixed subset. Cells report the two-judge mean and sample-bootstrap 95% confidence interval. AA uses only the 200 GT-authentic samples; all other columns use all 1,200 samples.

Method	ER	FC	VEG	HA	CFP	AA
FakeVLM	3.29 [3.22, 3.36]	3.09 [3.02, 3.16]	2.86 [2.80, 2.93]	3.59 [3.53, 3.65]	4.34 [4.31, 4.37]	3.84 [3.62, 4.04]
SIDA	3.87 [3.81, 3.92]	3.74 [3.68, 3.80]	3.20 [3.14, 3.26]	4.11 [4.06, 4.15]	4.39 [4.36, 4.41]	4.66 [4.60, 4.72]
MedForge	4.07 [4.03, 4.11]	3.99 [3.95, 4.03]	3.59 [3.55, 3.64]	3.90 [3.86, 3.94]	4.15 [4.13, 4.17]	4.71 [4.67, 4.74]
BEAR-MED	**4.15 [4.10, 4.19]**	**4.06 [4.02, 4.11]**	**3.66 [3.61, 3.70]**	**4.21 [4.17, 4.25]**	**4.48 [4.46, 4.50]**	**4.73 [4.69, 4.76]**

Table S31: Reasoning quality on the common-correct subset. Cells report the two-judge mean and sample-bootstrap 95% confidence interval. AA uses the 149 GT-authentic common-correct samples; all other columns use 986 samples.

Method	ER	FC	VEG	HA	CFP	AA
FakeVLM	3.43 [3.36, 3.51]	3.42 [3.35, 3.49]	3.01 [2.94, 3.09]	3.65 [3.59, 3.71]	4.37 [4.35, 4.40]	4.71 [4.66, 4.75]
SIDA	3.93 [3.87, 3.98]	3.89 [3.84, 3.95]	3.25 [3.18, 3.32]	4.09 [4.05, 4.14]	4.40 [4.37, 4.42]	4.72 [4.68, 4.77]
MedForge	4.07 [4.03, 4.12]	4.00 [3.96, 4.04]	3.60 [3.55, 3.65]	3.89 [3.84, 3.93]	4.14 [4.12, 4.17]	4.71 [4.66, 4.75]
BEAR-MED	**4.14 [4.09, 4.19]**	**4.07 [4.02, 4.11]**	**3.66 [3.61, 3.71]**	**4.18 [4.14, 4.23]**	**4.48 [4.45, 4.50]**	**4.74 [4.70, 4.78]**

Table S32: Paired full-subset differences, defined as BEAR-MED minus each baseline. Cells report mean difference and 95% sample-bootstrap interval.

Baseline	ER	FC	VEG	HA	CFP	AA
FakeVLM	.858 [.792, .923]	.977 [.905, 1.055]	.798 [.732, .861]	.621 [.563, .680]	.137 [.103, .172]	.885 [.675, 1.088]
SIDA	.281 [.229, .335]	.328 [.269, .388]	.458 [.399, .518]	.098 [.058, .140]	.088 [.058, .118]	.065 [.003, .140]
MedForge	.075 [.038, .110]	.076 [.037, .113]	.065 [.021, .105]	.310 [.268, .350]	.328 [.302, .351]	.023 [−.020, .068]

The paired intervals are positive for every non-AA comparison. The only interval crossing zero is BEAR-MED versus MedForge on AA; accordingly, the main paper treats their authenticity anchoring as statistically unresolved. The common-correct ranking is unchanged, which separates rationale quality from the direct effect of binary verdict errors.

H.4 JUDGE AGREEMENT AND CONSTRUCT DIAGNOSTICS

Table S33 compares the two judges' scores; Table S34 reports correlations between the six dimensions.

Table S33: Agreement between GPT-5.6 Sol and Gemini 3.1 Pro Preview.

Metric	n	Exact	Within 1	Mean $\lvert\Delta\rvert$
ER	4,800	.483	.953	.570
FC	4,800	.506	.951	.545
VEG	4,800	.425	.893	.691
HA	4,800	.435	.937	.631
CFP	4,800	.522	.983	.496
AA	800	.453	.983	.569

Table S34: Spearman correlations between two-judge mean scores on the full fixed subset. AA correlations use the 800 authentic method–sample items; other pairs use all 4,800 items.

	ER	FC	VEG	HA	CFP	AA
ER	1.000	.884	.820	.703	.470	.895
FC	.884	1.000	.782	.712	.478	.882
VEG	.820	.782	1.000	.630	.413	.756
HA	.703	.712	.630	1.000	.552	.772
CFP	.470	.478	.413	.552	1.000	.680
AA	.895	.882	.756	.772	.680	1.000

Within-one agreement ranges from .893 to .983 and exceeds exact agreement. VEG has the lowest within-one agreement and largest mean absolute difference. ER–FC ($\rho = .884$) and ER–VEG ($\rho = .820$) are strongly correlated, but the rubric distinguishes evidence relevance, visibility, and inference validity. Correlations diagnose the protocol rather than establish interchangeability. AA correlations use an authentic-only population and are not directly comparable with the core five-dimensional matrix. The audit package preserves frozen largest-disagreement records for qualitative adjudication without excluding them from aggregates.

I IMAGE-QUALITY DEGRADATION PROTOCOL AND COMPLETE RESULTS

The experiment fixes one class-balanced 10K test subset (5,000 authentic and 5,000 forged images) across all conditions. Its forged half has 1,000 examples per mechanism, so localization uses 4,000 local forgeries. Deterministic sampling is stratified by imaging category and generator family. BEAR-MED, MedForge, Mesorch, and TruFor run on clean images and 24 degraded copies, producing one million predictions.

The full sweep contains JPEG $Q \in \{100, 90, 80, 70, 60, 50\}$; Gaussian blur uses $\sigma \in \{0.50, 0.75, 1.00, 1.25, 1.50, 1.75\}$; Gaussian noise uses $\sigma \in \{3, 7, 11, 15, 19, 23\}$ in 8-bit intensity units; and downscaling uses factors $\{.95, .90, .80, .70, .60, .50\}$ followed by reconstruction to the canonical input size. A manifest fixes image identity, transform parameters, checksum, and GT geometry; all methods read identical cached images with boxes in canonical coordinates. No condition-specific training, calibration, thresholding, or checkpoint selection is performed.

The main-paper figure shows five representative levels per degradation; Tables S35, S36, S37, and S38 give the full JPEG, blur, noise, and downscale sweeps, including AUC. JPEG and blur progressively reduce strict localization, while noise is the strongest common failure mode. At downscale factor .50, BEAR-MED and Mesorch retain B@.75 of .706 and .703; at noise $\sigma = 23$, every method falls to .143 or below. The results support robustness to moderate degradation, not invariance to arbitrary corruption.

Table S35: Complete jpeg quality results on the fixed 10K degradation subset. Localization uses the 4,000 local-forgery rows.

	BEAR-Med			MedForge			Mesorch			TruFor		
quality Q	F1	AUC	B@.75	F1	AUC	B@.75	F1	AUC	B@.75	F1	AUC	B@.75
Clean	0.993	0.988	0.845	0.991	0.993	0.785	0.943	0.984	0.792	0.827	0.915	0.490
100	0.992	0.988	0.831	0.990	0.991	0.772	0.941	0.981	0.779	0.822	0.910	0.471
90	0.952	0.960	0.745	0.976	0.976	0.709	0.900	0.963	0.703	0.804	0.894	0.434
80	0.922	0.941	0.653	0.956	0.956	0.654	0.865	0.943	0.616	0.796	0.889	0.413
70	0.893	0.926	0.573	0.930	0.926	0.598	0.842	0.922	0.534	0.788	0.877	0.388
60	0.873	0.911	0.496	0.908	0.899	0.544	0.827	0.902	0.461	0.773	0.863	0.364
50	0.843	0.883	0.448	0.889	0.882	0.498	0.813	0.886	0.401	0.760	0.849	0.347

Table S36: Complete gaussian blur σ results on the fixed 10K degradation subset. Localization uses the 4,000 local-forgery rows.

	BEAR-Med			MedForge			Mesorch			TruFor		
blur σ	F1	AUC	B@.75	F1	AUC	B@.75	F1	AUC	B@.75	F1	AUC	B@.75
Clean	0.993	0.988	0.845	0.991	0.993	0.785	0.943	0.984	0.792	0.827	0.915	0.490
.50	0.964	0.974	0.824	0.984	0.985	0.765	0.940	0.981	0.788	0.822	0.910	0.483
.75	0.917	0.950	0.748	0.944	0.944	0.697	0.925	0.977	0.752	0.813	0.902	0.451
1.0	0.850	0.915	0.668	0.901	0.904	0.638	0.916	0.973	0.695	0.803	0.890	0.413
1.25	0.794	0.873	0.579	0.856	0.863	0.559	0.900	0.968	0.633	0.789	0.877	0.378
1.50	0.763	0.835	0.491	0.826	0.833	0.495	0.886	0.964	0.584	0.775	0.864	0.345
1.75	0.741	0.806	0.406	0.805	0.816	0.436	0.872	0.957	0.522	0.766	0.852	0.312

Table S37: Complete gaussian noise σ results on the fixed 10K degradation subset. Localization uses the 4,000 local-forgery rows.

	BEAR-Med			MedForge			Mesorch			TruFor		
noise σ	F1	AUC	B@.75	F1	AUC	B@.75	F1	AUC	B@.75	F1	AUC	B@.75
Clean	0.993	0.988	0.845	0.991	0.993	0.785	0.943	0.984	0.792	0.827	0.915	0.490
3	0.899	0.907	0.601	0.945	0.945	0.639	0.790	0.897	0.594	0.759	0.850	0.383
7	0.797	0.853	0.362	0.808	0.825	0.408	0.732	0.822	0.366	0.706	0.782	0.289
11	0.740	0.785	0.261	0.680	0.745	0.293	0.701	0.782	0.247	0.670	0.737	0.233
15	0.690	0.719	0.201	0.568	0.702	0.220	0.684	0.765	0.175	0.643	0.705	0.183
19	0.658	0.667	0.172	0.493	0.679	0.173	0.663	0.743	0.134	0.619	0.679	0.144
23	0.642	0.629	0.143	0.428	0.654	0.142	0.648	0.721	0.106	0.609	0.659	0.123

Table S38: Complete downscale factor results on the fixed 10K degradation subset. Localization uses the 4,000 local-forgery rows.

	BEAR-Med			MedForge			Mesorch			TruFor		
scale factor	F1	AUC	B@.75	F1	AUC	B@.75	F1	AUC	B@.75	F1	AUC	B@.75
Clean	0.993	0.988	0.845	0.991	0.993	0.785	0.943	0.984	0.792	0.827	0.915	0.490
.95	0.979	0.979	0.813	0.985	0.985	0.759	0.938	0.981	0.777	0.823	0.912	0.478
.90	0.971	0.975	0.816	0.982	0.982	0.756	0.935	0.980	0.780	0.824	0.913	0.479
.80	0.960	0.968	0.795	0.973	0.973	0.741	0.933	0.979	0.774	0.825	0.913	0.474
.70	0.943	0.956	0.780	0.964	0.964	0.721	0.926	0.976	0.757	0.821	0.910	0.471
.60	0.919	0.946	0.753	0.948	0.944	0.707	0.915	0.972	0.738	0.819	0.909	0.457
.50	0.853	0.912	0.706	0.904	0.891	0.673	0.867	0.948	0.703	0.812	0.902	0.426

J LIMITATIONS

Canonical resizing reduces format shortcuts but can alter native aspect ratio; the same transform is applied to all classes. Expert responses are learned evidence cues rather than causal proof. Uncommon authentic pathology can activate the medical-normality expert, and a fluent policy can still produce an unsupported narrative. Eligibility filtering, explicit FPR/FBR reporting, blinded rationale scoring, and automatic audits measure but do not eliminate these risks.

K QUALITATIVE EVIDENCE AND FAILURE ANALYSIS

This atlas presents 50 curated cases spanning dataset construction, expert evidence, and model behavior. Candidates from frozen records undergo visual inspection for plausible anatomy, acquisition texture, and edits; conspicuously artificial cases are excluded. Selection favors readable examples rather than random sampling, so the atlas does not estimate behavior prevalence. Images, predictions, rationales, scores, and audit labels are preserved without repair. Cyan denotes GT; orange dashed boxes denote predictions. Panel identifiers link to the selection manifest.

Dataset coverage. Figures S1–S4 contain 24 test images: one authentic, CM, SP, INS, REM, and SYN example in each of the four imaging categories. The optical examples include fundus, dermoscopy, and non-graphic endoscopy. These panels show the edited image and its annotation, not the model prediction.

Expert responses. Figure S5 uses four training examples, one per category, to illustrate overlapping local cues, broader generation cues, and authentic tissue that activates the medical-normality expert. These are qualitative response patterns, not mechanism-specific decision rules.

Localization and reasoning comparisons. Figures S6–S7 compare four local forgeries, pairing CM, SP, INS, and REM with C1–C4. All columns use the same image and detail crop; missing predictions remain absent. The four reasoning comparisons in Figures S8–S11 come from the common-correct subset: FakeVLM, SIDA, MedForge, and BEAR-MED all have correct detection verdicts. They cover authentic, SP, INS, and SYN images across C1, C3, C2, and C4, respectively. Each panel preserves the full rationale supplied to the judges and reports their mean scores. These comparisons isolate explanation quality when detection agrees. Generated explanations remain model outputs, not clinical findings.

Failure modes. Figures S12–S14 contain six BEAR-MED cases: a missed edit, displaced box, over-large box, authentic false alarm, local box for global synthesis, and fluent but weakly grounded reasoning. In the last case, the predicted box overlaps GT well, yet both judges assign VEG a score of 2. Correct localization therefore does not guarantee that the rationale describes the right visual evidence.

Degradation and external transfer. Figures S15–S18 show four cached clean–degraded triplets with frozen predictions, one per degradation family and imaging category. Figures S19–S20 show four MedForge-90K cases without target-dataset adaptation: two INS and two REM, spanning chest radiography, brain imaging, and fundus. All four cases have recorded CONTAINED audit outcomes.

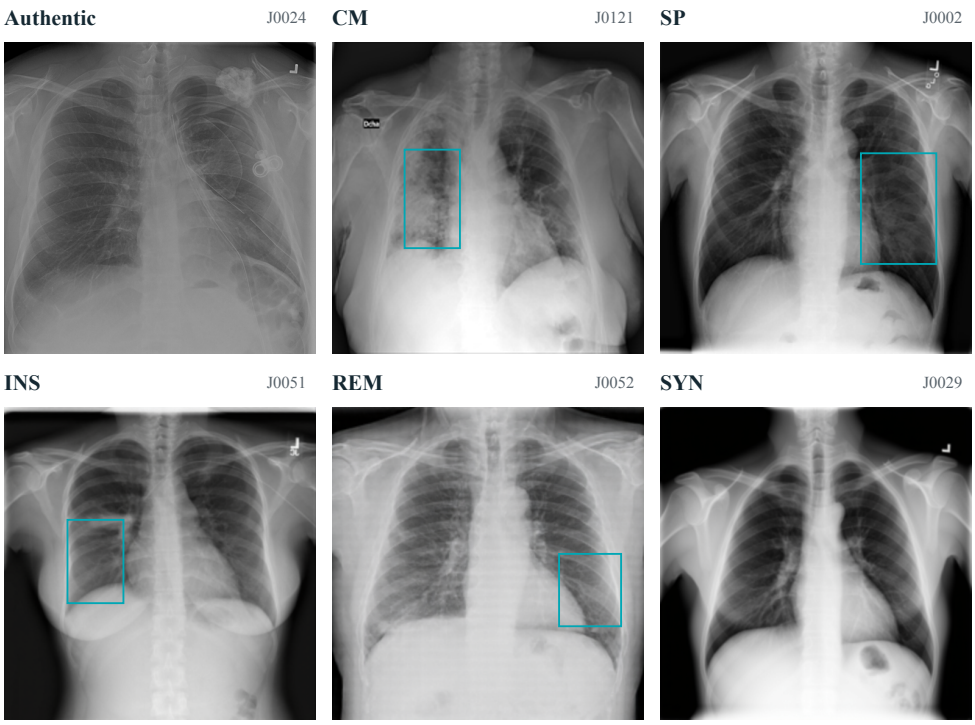


Figure S1: Planar-radiography examples of authentic images and all five forgery mechanisms. Cyan boxes mark annotated local edits.

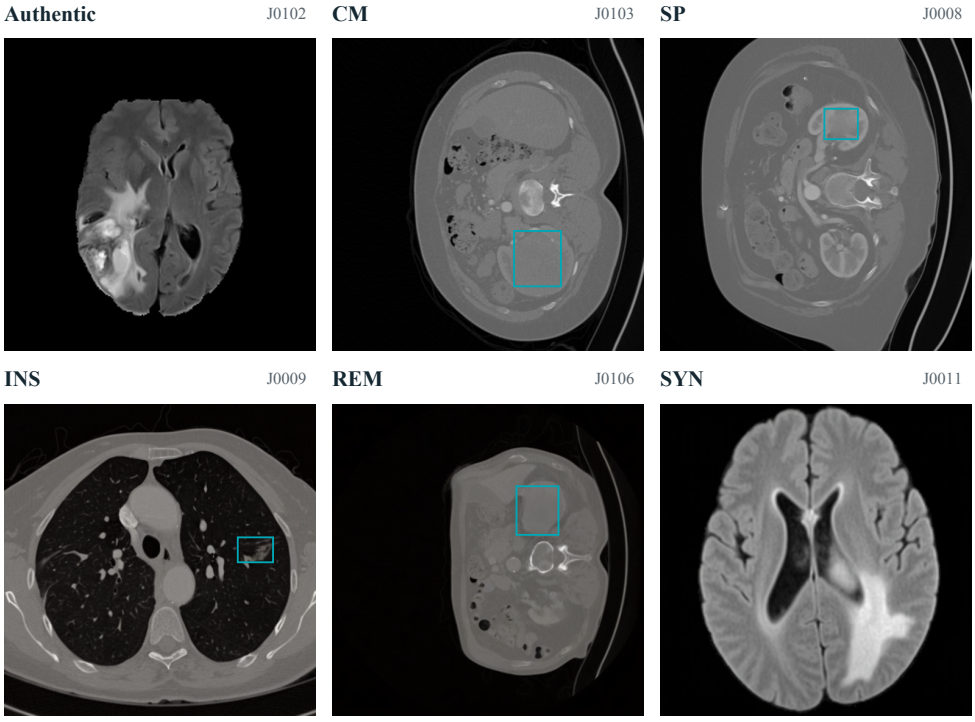


Figure S2: CT/MRI examples of authentic images and all five forgery mechanisms. Cyan boxes mark annotated local edits.

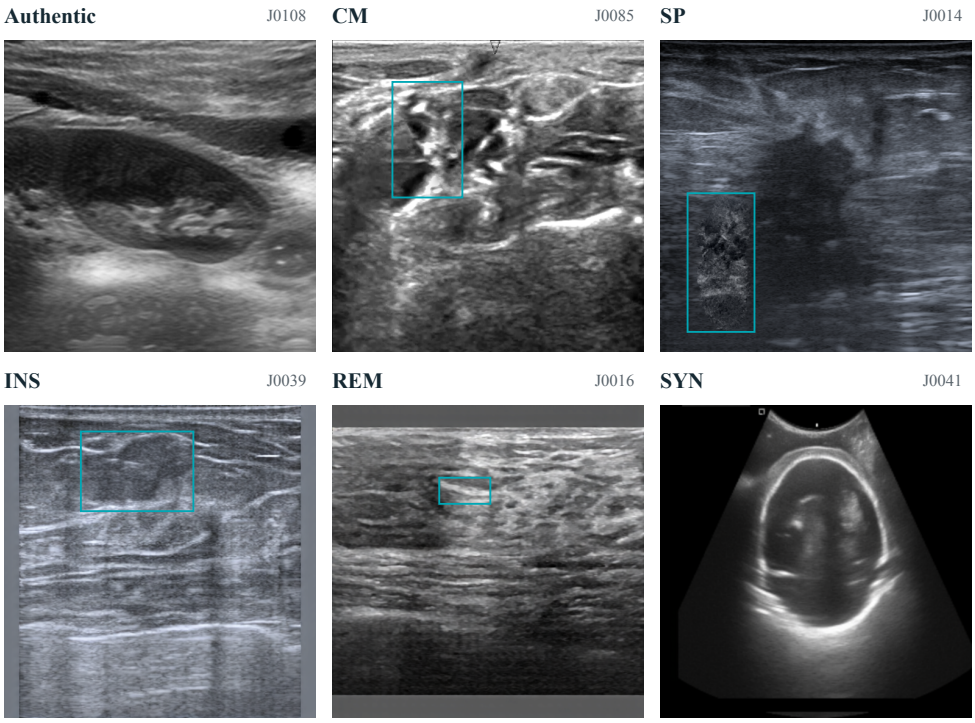


Figure S3: Ultrasound examples of authentic images and all five forgery mechanisms. Cyan boxes mark annotated local edits.

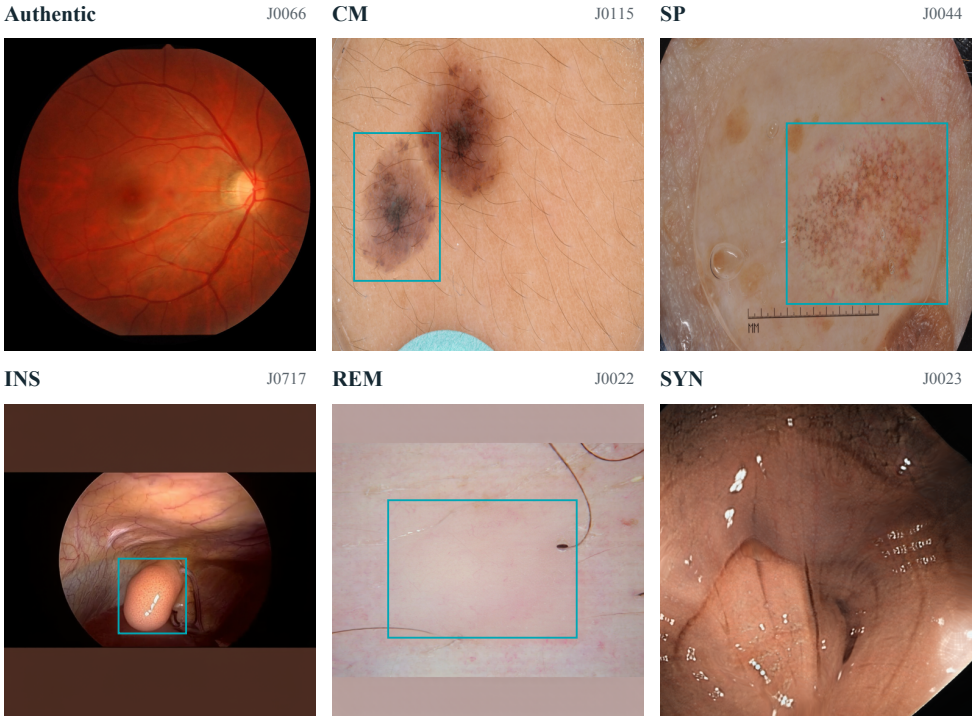


Figure S4: Optical examples spanning fundus, dermoscopy, and endoscopy. Each local edit has an annotated box; authentic and fully synthetic images have no local target.

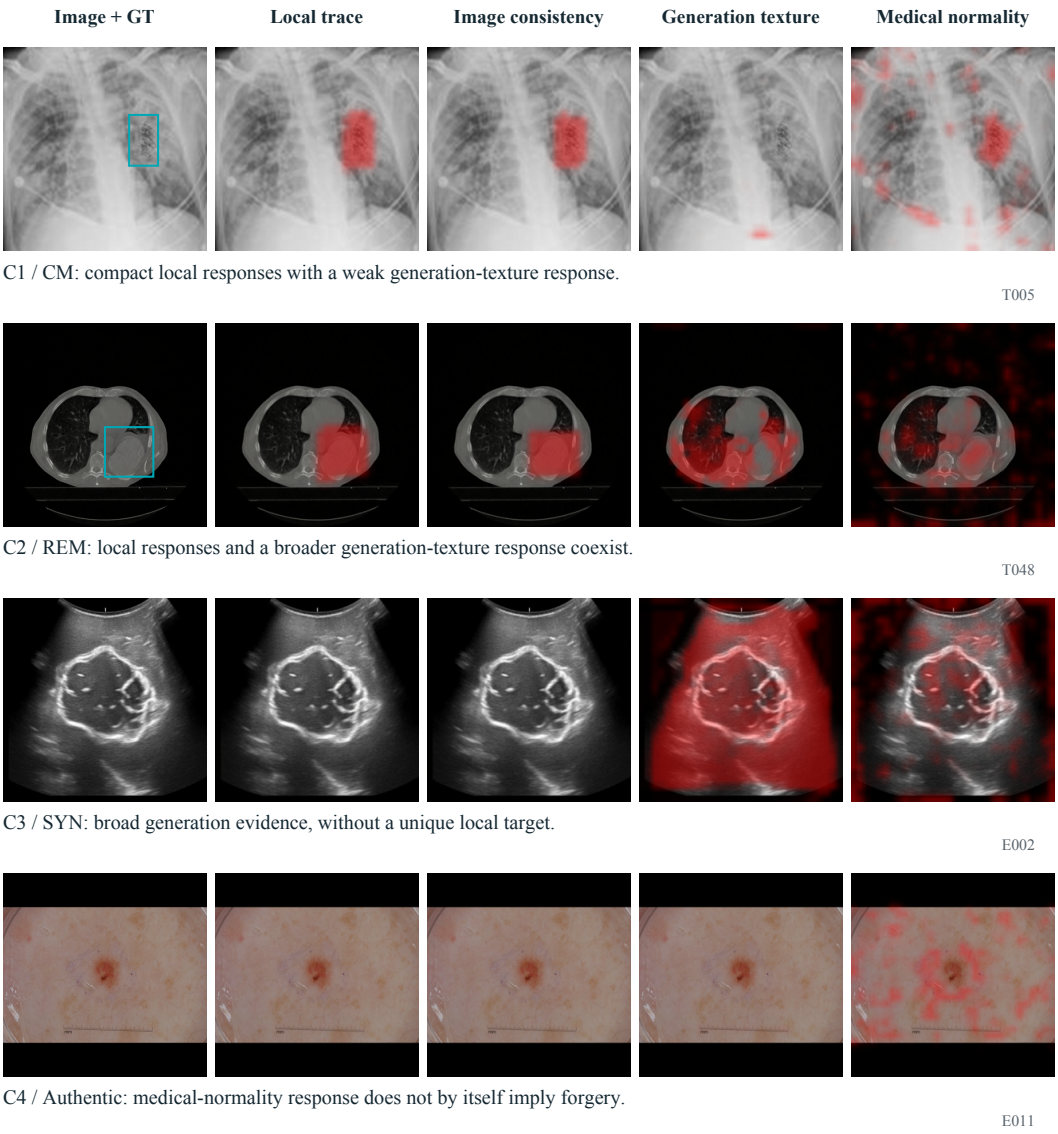


Figure S5: Complementary frozen-expert responses on four training examples. Cached overlays share an absolute response scale and fixed opacity. Medical-normality evidence can also be strong on an authentic image.

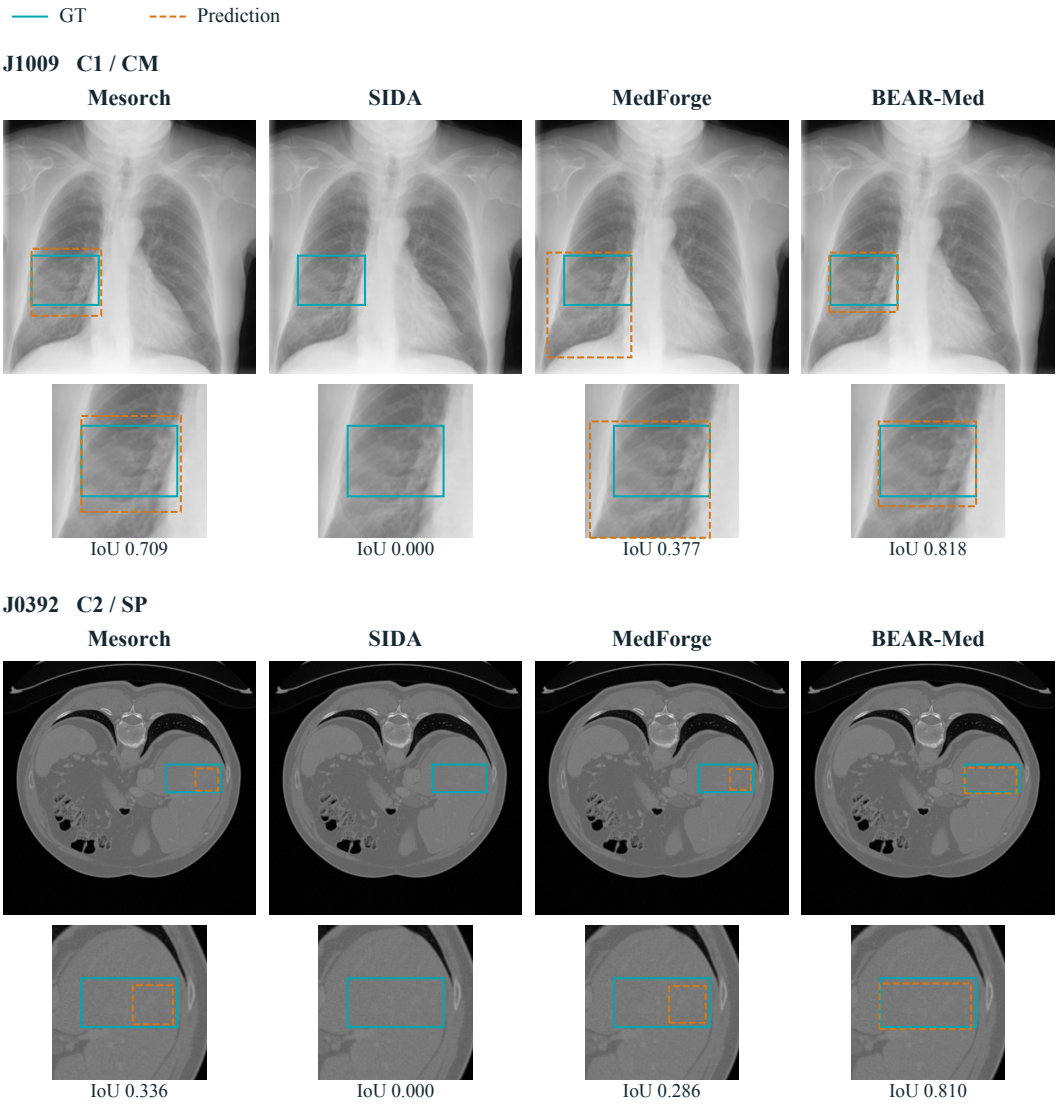


Figure S6: Localization comparison on C1/CM and C2/SP test images. Each column shows the full image and an identical GT-centered detail crop. IoU is computed from the frozen predicted and GT boxes.

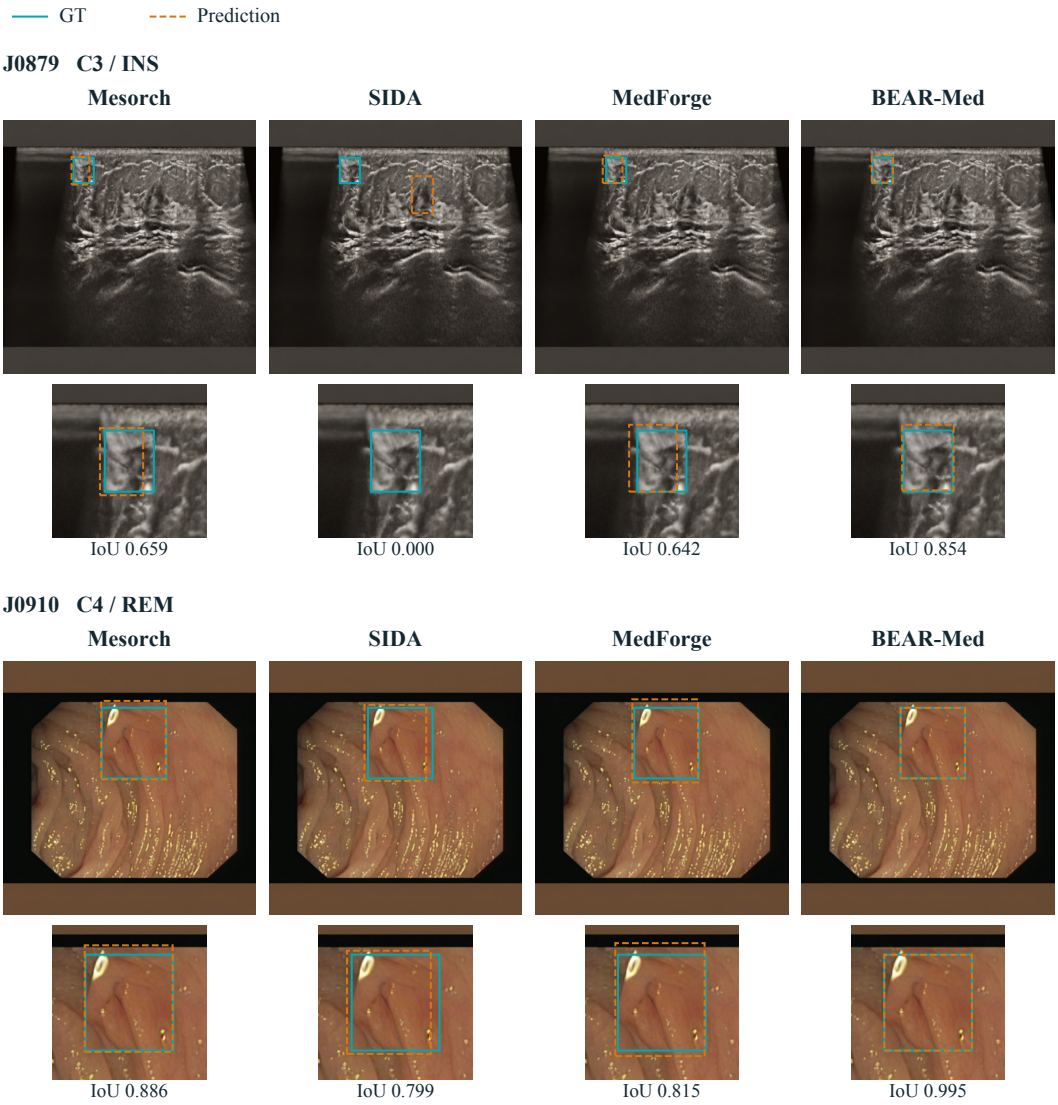
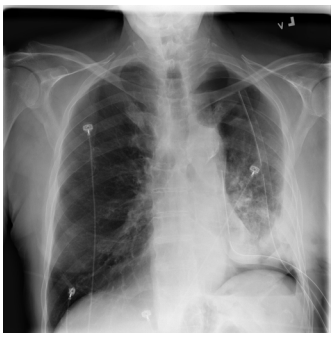


Figure S7: Localization comparison on C3/INS and C4/REM test images. The C4 case also shows close predictions from competing methods, rather than requiring every baseline to fail.



J0816 C1 / Authentic

	ER	FC	VEG	HA	CFP	AA
FakeVLM	4.5	4.5	4.0	4.5	4.5	4.5
SIDA	4.5	4.5	3.5	4.5	4.5	4.5
MedForge	4.5	4.5	4.5	4.5	4.5	4.5
BEAR-Med	5.0	5.0	5.0	5.0	5.0	5.0

Mean scores of the two judges (1-5).
AA applies to authentic images only.
All four detection verdicts are correct.

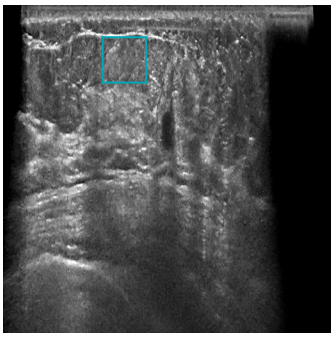
FakeVLM. real. The chest radiograph demonstrates continuous anatomical structures, including intact rib contours and consistent lung markings across both hemithoraces. The diffuse opacity in the left lower zone exhibits plausible density transitions and respects the cardiac silhouette without abrupt truncation or sticker-like borders. Global image texture, noise distribution, and contrast remain uniform throughout the field of view, showing no signs of local manipulation or synthetic artifacts. These features collectively support the authenticity of the acquisition as a genuine clinical projection X-ray.

SIDA. The chest radiograph demonstrates continuous anatomical structures, including intact rib contours, clear lung markings, and a coherent cardiac silhouette without abrupt truncation or boundary artifacts. The image exhibits consistent global contrast and detector noise patterns typical of authentic projection X-ray acquisition, with no evidence of local texture smoothing or inconsistent exposure. Medical devices, such as the right-sided chest tube and ECG leads, display plausible positioning and natural interaction with underlying anatomy. These features collectively support the authenticity of the image as a genuine clinical radiograph.

MedForge. The chest radiograph displays continuous anatomical structures, including intact rib contours and consistent lung markings across both fields. Medical devices, such as ECG leads and a left-sided chest tube, are visible with plausible positioning. The medical devices exhibit appropriate radiographic density without edge artifacts or texture discontinuities, which would otherwise suggest manipulation. The global noise pattern and contrast remain uniform throughout the image, adhering to expected projection X-ray physics. These coherent structural and textural features indicate that the image lacks the inconsistencies typically associated with synthetic generation or editing. The coherent anatomical and textural features support the classification of this as a real image.

BEAR-Med. The chest radiograph demonstrates continuous anatomical structures with preserved lung markings and consistent rib alignment across both hemithoraces. Image texture exhibits uniform detector noise and natural stochastic variation without evidence of local smoothing, abrupt contrast shifts, or sticker-like borders. The visible medical devices and pathological opacities integrate seamlessly with the surrounding tissue density and global exposure characteristics. These features collectively support an authentic acquisition with no observable forensic inconsistencies or manipulation artifacts.

Figure S8: Common-correct reasoning comparison on an authentic radiograph. Full original rationales and mean scores of the two judges are shown; all four detection verdicts are correct.



J1070 C3 / SP

	ER	FC	VEG	HA	CFP	AA
FakeVLM	2.0	2.0	1.5	2.5	4.0	-
SIDA	3.0	3.0	1.5	4.0	4.0	-
MedForge	3.5	3.5	1.5	4.0	3.5	-
BEAR-Med	4.5	4.5	4.5	4.5	5.0	-

Mean scores of the two judges (1-5).
AA applies to authentic images only.
All four detection verdicts are correct.

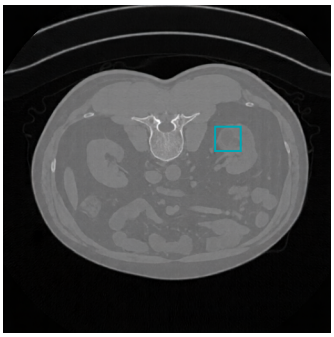
FakeVLM. Although the image displays a plausible breast ultrasound appearance with a central hypoechoic focus and posterior shadowing, the global texture exhibits synthetic characteristics inconsistent with authentic acquisition. The speckle pattern appears uniformly generated and lacks the natural stochastic variation and depth-dependent gain attenuation expected in real sonographic data. While a benign cyst or fibroadenoma could explain the focal finding, the pervasive textural uniformity across the entire field of view suggests full-image synthesis rather than a genuine clinical scan. Consequently, the absence of authentic acoustic physics artifacts outweighs the superficial anatomical plausibility.

SIDA. The central hypoechoic region exhibits subtle textural inconsistencies and boundary irregularities that disrupt the natural continuity of the surrounding fibroglandular tissue planes. While a genuine cyst or solid mass could present as a focal dark area, authentic lesions typically demonstrate coherent posterior acoustic features and seamless integration with adjacent speckle patterns. Here, the lesion's margins appear artificially distinct without corresponding beam interaction artifacts, suggesting a localized splicing manipulation rather than a biologically plausible pathology. This structural discontinuity weighs against a benign anatomical explanation despite the overall ultrasound appearance.

MedForge. The image displays a central anechoic structure with abnormally sharp, well-defined margins and uniform internal darkness. This finding is situated within surrounding heterogeneous parenchymal speckle. The structure lacks the subtle boundary fuzziness and posterior acoustic enhancement typically seen in authentic fluid-filled lesions. The abrupt textural transition between the lesion and the surrounding tissue suggests a spliced insertion rather than natural tissue interaction. Furthermore, the absence of expected beam attenuation artifacts indicates that the region was likely composited from a separate source with inconsistent acquisition physics. The visual inconsistencies in texture and acoustic artifacts indicate that the image is a deepfake.

BEAR-Med. The upper-central tissue region exhibits subtle textural inconsistencies and boundary irregularities that disrupt the natural continuity of the surrounding fibroglandular architecture. While a benign focal lesion could theoretically present as a discrete area, the lack of coherent posterior acoustic features and the abrupt transition in speckle statistics argue against a genuine anatomical finding. This localized discontinuity suggests an artificial insertion rather than a naturally occurring pathology, as authentic tissue planes typically maintain gradual integration with adjacent structures.

Figure S9: Common-correct reasoning comparison on an ultrasound SP image. The outputs differ in whether they describe a local edit or global synthesis and where they place the supporting cues.



J0297 C2 / INS

	ER	FC	VEG	HA	CFP	AA
FakeVLM	2.0	2.0	2.0	2.0	4.0	-
SIDA	2.0	2.0	2.0	2.0	4.0	-
MedForge	3.0	2.0	3.0	2.5	4.0	-
BEAR-Med	5.0	4.5	4.5	5.0	4.5	-

Mean scores of the two judges (1-5).
AA applies to authentic images only.
All four detection verdicts are correct.

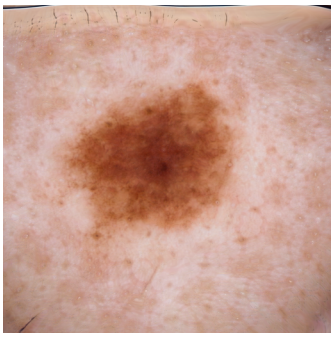
FakeVLM. cipline. This is a fake image. Although the abdominal anatomy appears structurally plausible with correct organ positioning, the image exhibits a pervasive, uniform texture that lacks the stochastic noise and subtle reconstruction artifacts characteristic of authentic CT acquisitions. This global smoothness suggests synthetic generation rather than natural photon statistics, as real scans typically display heterogeneous grain patterns even in homogeneous soft tissue regions. Consequently, the absence of genuine acquisition physics across the entire field of view outweighs the anatomical correctness, indicating the image is likely fully synthesized.

SIDA. The abdominal CT exhibits diffuse, generation-like texture artifacts and global statistical deviations from authentic medical imaging patterns across the entire field of view. While a plausible benign alternative for the subtle irregularity in the left upper quadrant could be partial volume averaging or normal bowel content, the pervasive synthetic noise structure contradicts this interpretation. The lack of consistent acquisition physics and coherent tissue boundaries throughout the scan confirms that the image is not a genuine clinical acquisition but rather a synthesized forgery lacking realistic anatomical fidelity.

MedForge. The abdominal CT image displays a diffuse, generation-like texture inconsistency across the entire field of view, deviating from the stochastic noise pattern expected in authentic tomographic acquisitions. A focal density variation is visible in the left upper quadrant near the renal region. The pervasive textural uniformity and global acquisition anomalies suggest the image is a full synthesis rather than a genuine scan with a localized edit. Although a focal density variation is present in the left upper quadrant, this region lacks the sharp boundary artifacts or local structural discontinuities typically associated with isolated lesion insertion. The absence of these specific local manipulation traces, combined with the inconsistent texture, supports the hypothesis of a deepfake. The image is a deepfake.

BEAR-Med. A focal hypodense lesion is visible in the left upper quadrant, appearing as a distinct area of altered density within the renal parenchyma. While such findings can represent authentic cysts or tumors, this region exhibits subtle boundary inconsistencies and lacks the expected mass effect or tissue displacement on adjacent structures. The texture within the lesion appears unnaturally uniform compared to the stochastic noise pattern of the surrounding kidney tissue, suggesting an artificial insertion rather than a biological pathology. These local structural discontinuities weigh against a benign etiology, indicating the finding is likely a synthesized artifact.

Figure S10: Common-correct reasoning comparison on a CT INS image. Original text, including textual irregularities, is preserved. AA is not applicable to forged images.



J1103 C4 / SYN

	ER	FC	VEG	HA	CFP	AA
FakeVLM	3.5	4.0	3.0	4.0	4.5	-
SIDA	3.5	4.0	3.5	4.5	5.0	-
MedForge	4.0	4.5	3.5	4.0	4.5	-
BEAR-Med	4.0	5.0	4.0	5.0	5.0	-

Mean scores of the two judges (1-5).
AA applies to authentic images only.
All four detection verdicts are correct.

FakeVLM. Although the central pigmented lesion displays a plausible reticular pattern consistent with benign melanocytic nevi, the surrounding skin texture exhibits unnatural uniformity and oversmoothing characteristic of synthetic generation. Authentic dermoscopy typically presents stochastic micro-texture and variable illumination across the field, whereas this image shows a globally consistent, plastic-like surface lacking natural sensor noise or pore detail. The absence of realistic skin topography and the overly regular background pattern weigh against an authentic acquisition, indicating the entire image is likely synthesized rather than captured from a patient.

SIDA. Although the lesion displays a plausible pigment network and border fading consistent with benign dermoscopic patterns, the surrounding skin exhibits an unnaturally uniform texture lacking expected micro-relief and stochastic noise. This global oversmoothing suggests synthetic generation rather than authentic optical acquisition, as real skin typically retains fine surface irregularities and sensor noise even in out-of-focus areas. Consequently, the absence of natural textural variance across the entire field outweighs the superficial clinical plausibility of the central finding.

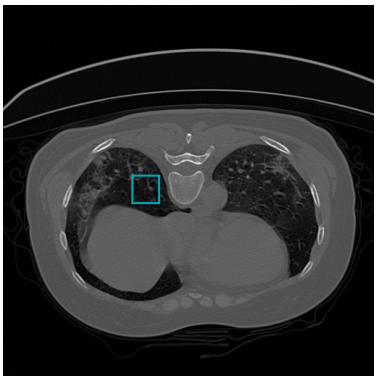
MedForge. The image displays a lesion with a pigment network and border fading, set against surrounding skin that exhibits an unnaturally uniform texture. While the lesion itself shows features consistent with benign dermoscopic patterns, the surrounding skin lacks the stochastic micro-relief and fine vellus hairs typical of authentic optical imaging. This global oversmoothing and the absence of natural sensor noise indicate that the entire field was synthetically generated rather than captured from real tissue. The pervasive lack of genuine acquisition artifacts across the background further supports the hypothesis of full synthesis. The pervasive lack of genuine acquisition artifacts and global oversmoothing confirms that the image is a deepfake.

BEAR-Med. Although the lesion displays a plausible pigment network and border fading consistent with benign dermoscopic patterns, the surrounding skin texture exhibits unnatural uniformity and oversmoothing that lacks authentic stochastic detail. This global textural inconsistency suggests synthetic generation rather than natural optical acquisition, as real skin typically retains fine micro-relief and sensor noise even in out-of-focus areas. Consequently, the absence of genuine tissue granularity across the entire field outweighs the superficial clinical plausibility of the central finding.

Figure S11: Common-correct reasoning comparison on a dermoscopy SYN image. All four methods return the correct verdict; the recorded explanations and scores show both shared observations and differences in specificity.

GT Prediction

Missed local edit



F012 C2 / SP

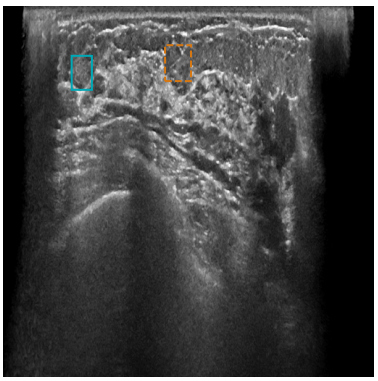
Original rationale

The axial chest CT demonstrates authentic acquisition characteristics, including consistent noise texture and appropriate Hounsfield unit density across lung parenchyma, mediastinum, and osseous structures. Anatomical boundaries are continuous and coherent, with no evidence of boundary discontinuities, smudging, or unnatural smoothing that would suggest manipulation. The vascular markings and tissue transitions follow expected physiological patterns without structural inconsistencies or impossible overlaps. These features collectively support the image's authenticity as a genuine clinical scan.

Prediction: authentic; box IoU: 0.000

The local SP edit is classified as authentic, so no box is returned.

Displaced box



F000 C3 / SP

Original rationale

The upper-central subcutaneous tissue exhibits a localized region of heterogeneous echotexture that disrupts the continuity of adjacent fibroglandular planes. While such irregularity could represent a benign focal lesion or artifact, the area lacks coherent posterior acoustic features and displays subtle boundary inconsistencies inconsistent with natural tissue interfaces. The speckle pattern within this zone appears statistically distinct from the surrounding parenchyma, suggesting an inserted texture rather than authentic anatomical variation. These structural discontinuities and textural mismatches weigh against a plausible benign etiology, indicating local manipulation.

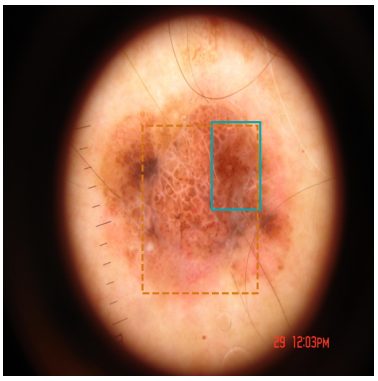
Prediction: forged; box IoU: 0.000

The image is correctly classified as forged, but the returned box misses the annotated SP region.

Figure S12: Missed-edit and displaced-box failures. The original rationales accompany the frozen predictions; cyan is GT and orange is the predicted box.

GT Prediction

Over-large box



F069 C4 / SP

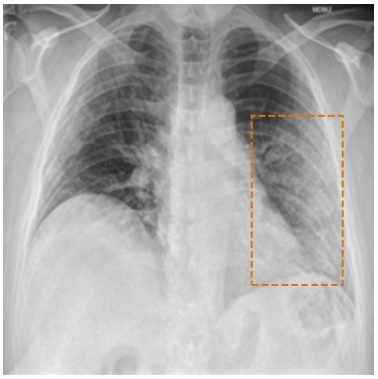
Original rationale

The central pigmented lesion exhibits diffuse textural inconsistencies and generation-like artifacts that disrupt the natural stochastic pattern expected in authentic dermoscopy. While irregular pigment networks can occur in benign nevi, the observed smudging and lack of coherent microstructures across the upper region suggest synthetic manipulation rather than biological variation. The abrupt transition between the lesion's core and the surrounding skin further indicates a local copy-move or synthesis artifact inconsistent with genuine tissue continuity.

Prediction: forged; box IoU: 0.195

The predicted box covers substantially more area than the annotated SP edit.

Authentic false alarm



F092 C1 / Authentic

Original rationale

The left mid-to-lower lung field exhibits a distinct rectangular region of altered texture and density that disrupts the natural continuity of pulmonary vascular markings. While this opacity could theoretically represent a dense consolidation or atelectasis, the sharply defined geometric borders and lack of associated volume loss signs, such as diaphragmatic elevation or mediastinal shift, argue against a genuine pathological process. Instead, the abrupt transition between the modified area and surrounding aerated lung suggests a localized manipulation where background structures were not seamlessly preserved. This structural inconsistency within the lung parenchyma indicates an artificial alteration rather than authentic clinical disease.

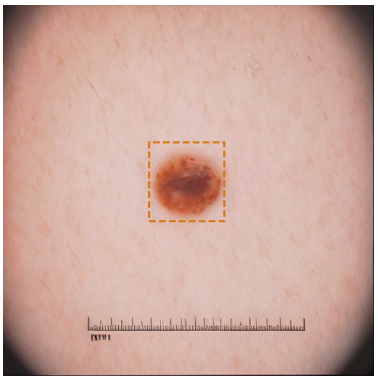
Prediction: forged; local box returned

An authentic image is classified as forged and assigned a local box.

Figure S13: Over-large-box and authentic-false-alarm failures. A plausible medical image can still trigger an incorrect spatial extent or forgery verdict.

GT Prediction

Local box for global synthesis



F093 C4 / SYN

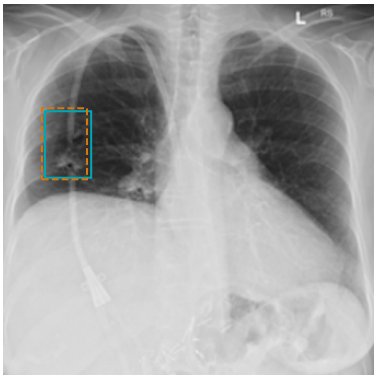
Original rationale

While the central pigmented lesion appears morphologically plausible for a dermoscopic examination, the surrounding skin exhibits an unnaturally uniform texture lacking the stochastic micro-relief and fine vellus hairs typical of authentic optical imaging. This pervasive oversmoothing suggests the background is synthetically generated or heavily processed rather than representing natural tissue variation. Although a benign alternative could attribute this appearance to aggressive noise reduction or focus falloff, the complete absence of sensor noise and consistent high-frequency detail across the entire field contradicts standard acquisition physics. Consequently, the global textural inconsistency outweighs the local clinical plausibility of the lesion itself.

Prediction: forged; local box returned

The SYN image is classified as forged, but the response incorrectly assigns a local box.

Fluent but weakly grounded reasoning



J0217 C1 / CM

Original rationale

The right upper lung field exhibits subtle textural inconsistencies and boundary irregularities that disrupt the natural continuity of pulmonary vascular markings. While a benign etiology like localized atelectasis or scarring could explain focal opacity, the absence of associated volume loss signs, such as fissure displacement or mediastinal shift, weighs against this authentic alternative. Furthermore, the region demonstrates internal texture uniformity that contrasts with the stochastic noise pattern of the surrounding aerated lung, suggesting a copy-move manipulation rather than genuine pathology. These structural discontinuities indicate that the area does not preserve the expected overlapping background anatomy consistent with projection radiography physics.

Prediction: forged; box IoU: 0.803

Both judges assign VEG = 2: professional wording does not establish correspondence to the edited region.

Figure S14: Scope and reasoning failures. Top: a local box is returned for SYN. Bottom: high geometric overlap coexists with weak visual grounding in the rationale.

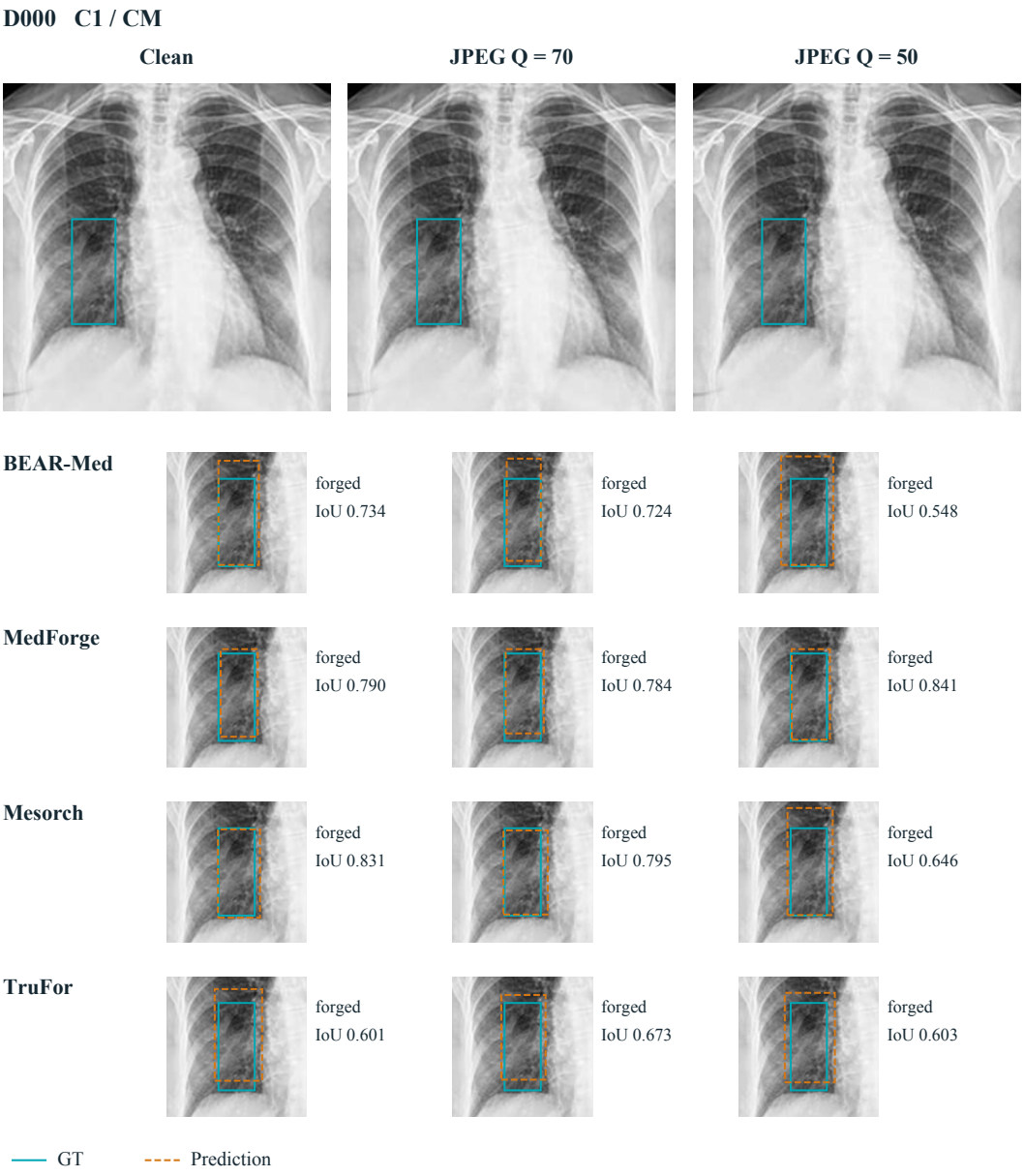


Figure S15: JPEG degradation of a radiographic CM example. Top: full images and GT. Below: matched detail crops with predictions, verdicts, and geometric box IoU.

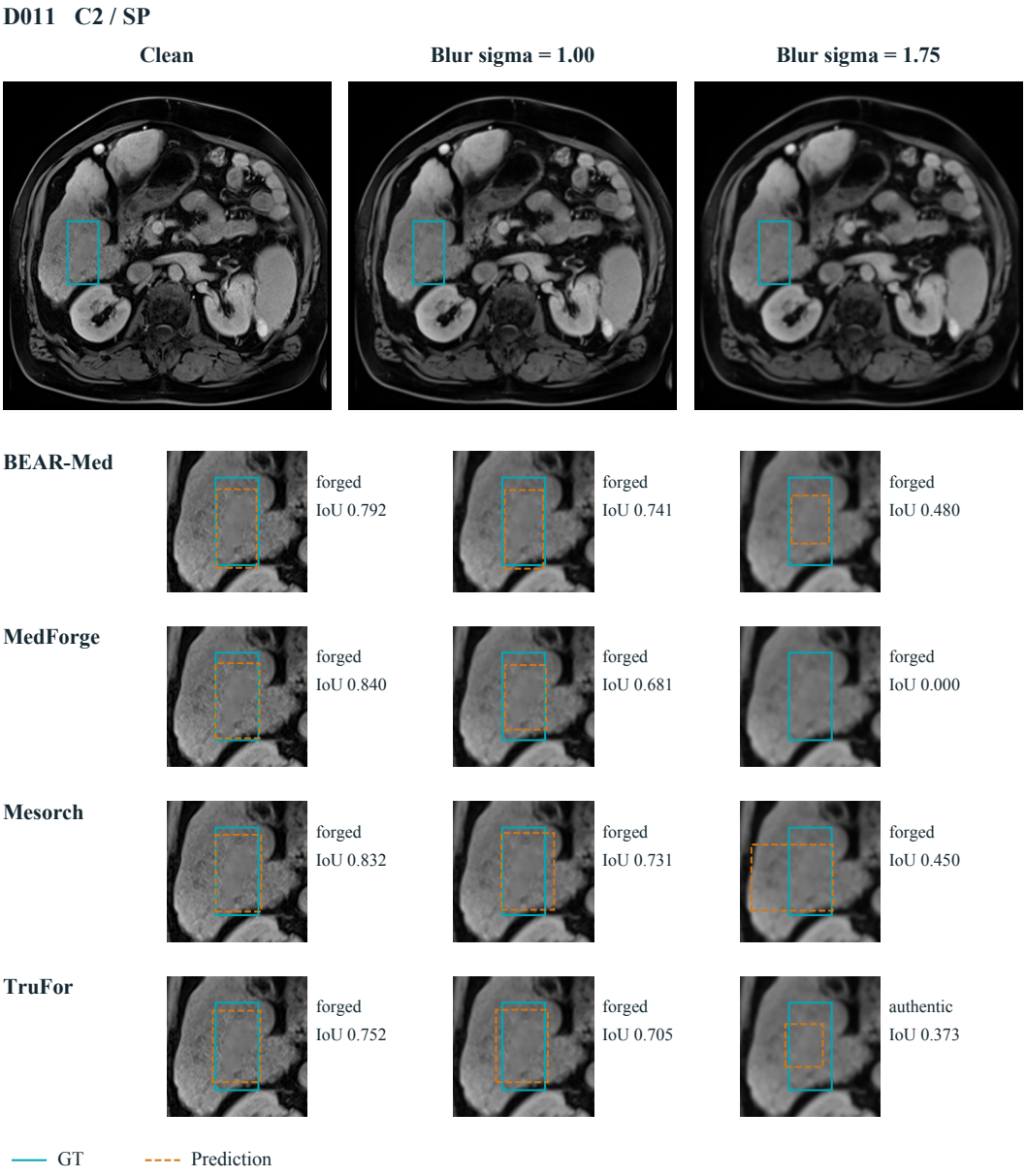


Figure S16: Gaussian-blur degradation of a CT SP example, using the same crop and display convention.

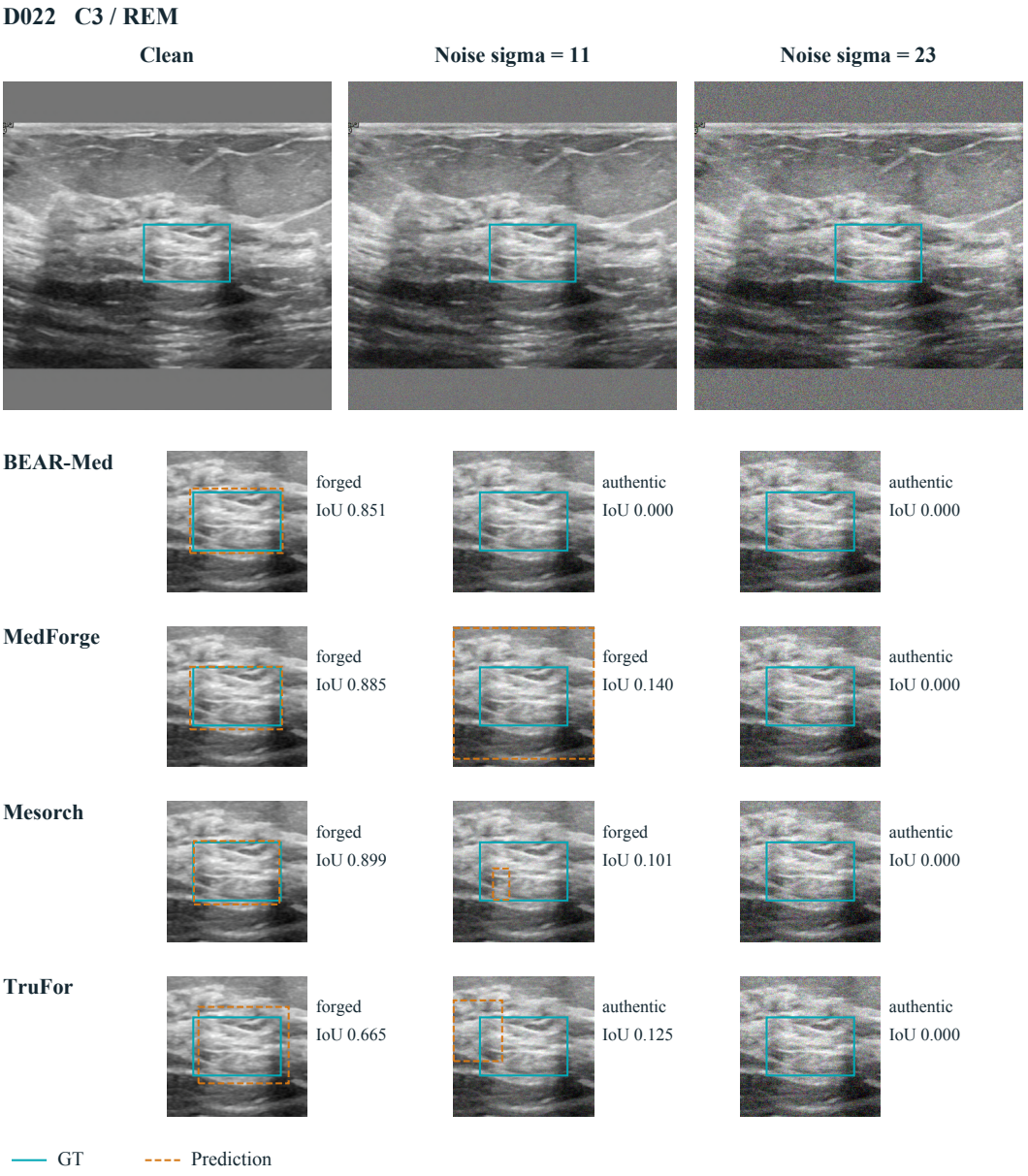


Figure S17: Gaussian-noise degradation of an ultrasound REM example. Both the images and predictions are reused from the completed evaluation.

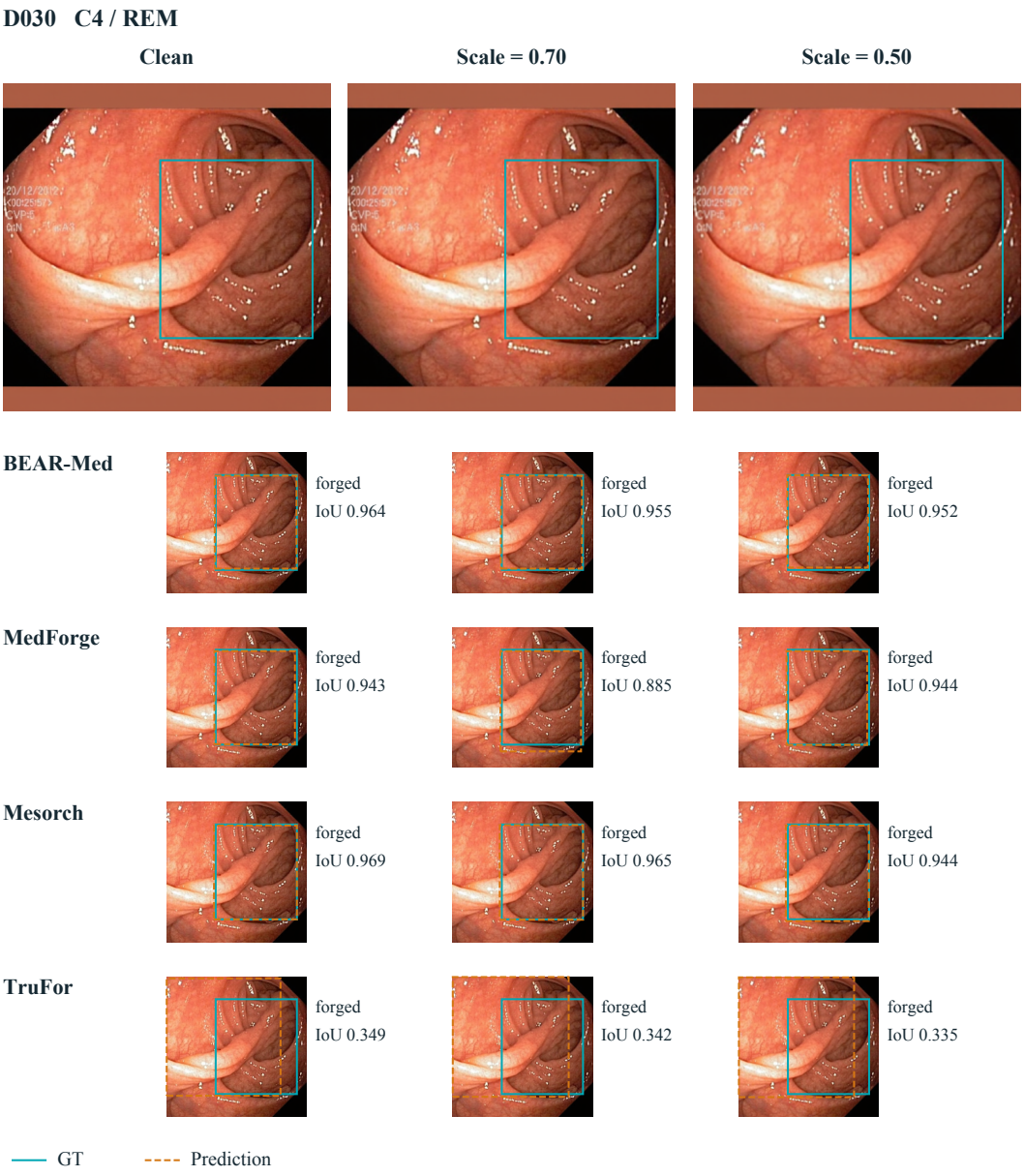
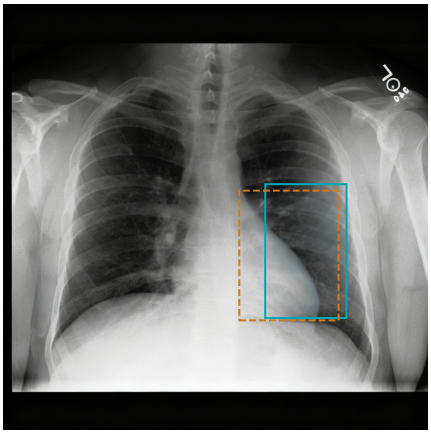


Figure S18: Downscale-and-reconstruction degradation of an endoscopy REM example. Each method is evaluated on the same three cached images.

GT Prediction

X071 chest-xray / INS



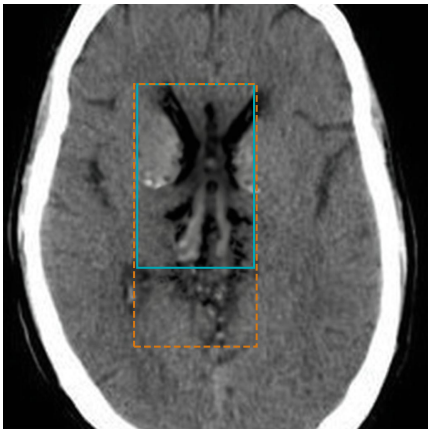
BEAR-Med rationale

The left mid-to-lower lung zone exhibits an unnatural, homogeneous opacity that lacks the stochastic texture and vascular markings expected in authentic radiography. While a large consolidation or effusion could clinically explain increased density, the region shows abrupt boundary transitions and fails to demonstrate plausible mass effect or anatomical displacement of adjacent structures. This structural discontinuity and uniform texture pattern are inconsistent with natural tissue overlap, indicating a localized manipulation rather than genuine pathology.

Prediction: forged; IoU 0.648

Recorded audit: CONTAINED

X065 Brain imaging / INS



BEAR-Med rationale

The central brain parenchyma exhibits diffuse textural smoothing and inconsistent noise characteristics compared to the surrounding cortical tissue, suggesting synthetic alteration rather than natural acquisition. While severe pathology could theoretically cause homogeneous signal changes, the absence of corresponding mass effect, ventricular distortion, or plausible anatomical boundaries contradicts a benign clinical explanation. The abrupt transition between the altered central region and the preserved peripheral anatomy further indicates a localized manipulation inconsistent with authentic tomographic imaging physics.

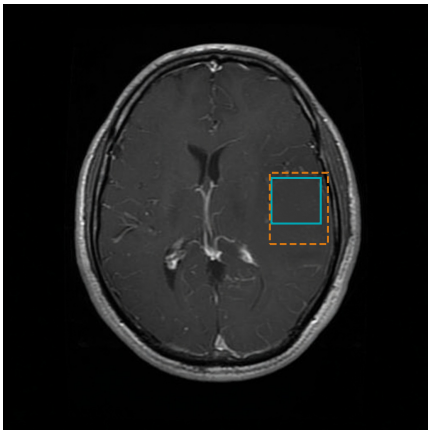
Prediction: forged; IoU 0.666

Recorded audit: CONTAINED

Figure S19: External INS examples from chest radiography and brain imaging. Cyan and orange show GT and predicted boxes; both pairs have recorded CONTAINED audit outcomes.

GT Prediction

X050 Brain imaging / REM



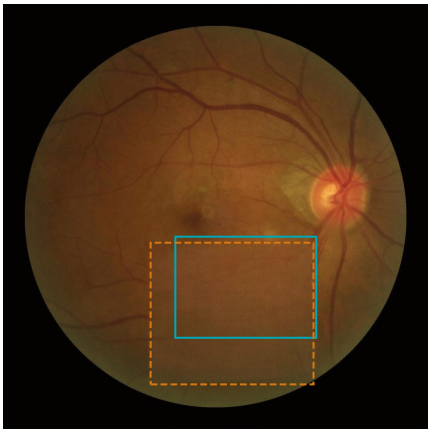
BEAR-Med rationale

The brain MRI exhibits diffuse generation-like texture artifacts and global signal inconsistencies that deviate from authentic acquisition patterns, particularly within the deep white matter and ventricular regions. While a benign alternative might attribute these features to severe motion artifact or low-resolution reconstruction, the uniform smoothing and lack of coherent anatomical boundaries across the entire slice contradict typical imaging physics. The absence of natural stochastic noise and the presence of structurally implausible tissue transitions indicate synthetic manipulation rather than a genuine clinical scan.

Prediction: forged; IoU 0.542

Recorded audit: CONTAINED

X100 fundus / REM



BEAR-Med rationale

The inferior temporal retina exhibits an unnatural, uniform haze that lacks the stochastic granular texture and vessel sharpness seen in the superior field. While this smooth appearance could theoretically result from severe media opacity or defocus, the abrupt transition to normal retinal detail without corresponding blur gradients suggests artificial smoothing rather than optical degradation. The absence of expected microvascular structures within this diffuse region further indicates a manipulation trace consistent with lesion removal or generative inpainting.

Prediction: forged; IoU 0.543

Recorded audit: CONTAINED

Figure S20: External REM examples from brain MRI and fundus imaging. The final checkpoint is used without target-dataset adaptation; original rationales and geometric overlap are shown.

REFERENCES

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. *arXiv preprint arXiv:2511.21631*, 2025.

Black Forest Labs. FLUX.1 Fill [dev]. Official model card, 2024. URL `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`.

Zhihui Chen, Kai He, Qingyuan Lei, Bin Pu, Jian Zhang, Yuling Xu, and Mengling Feng. MedForge: Interpretable medical deepfake detection via forgery-aware reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25672–25692. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1177.

Junlong Cheng, Bin Fu, Jin Ye, Guoan Wang, Tianbin Li, Haoyu Wang, Ruoyu Li, He Yao, Junren Cheng, Jingwen Li, Yanzhou Su, Min Zhu, and Junjun He. Interactive medical image segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20841–20851, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 852–863, 2021.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Shaked Perek, Ben Wiesel, Avihu Dekel, Nimrod Shabtay, and Eli Schwartz. Balanced thinking: Improving chain of thought training in vision language models. *arXiv preprint arXiv:2603.18656*, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, 2022.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, 2023.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. SeqTR: A simple yet universal network for visual grounding. In *European Conference on Computer Vision (ECCV)*, pp. 598–615. Springer, 2022. doi: 10.1007/978-3-031-19833-5_35.