LEARNING TEXT-CONDITIONED STATE-SPACE RECURRENCE FOR MULTIMODAL MEDICAL LOGICAL DEEPFAKE VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models can produce biomedical images that are physically plausible yet contradict their accompanying text. We call such inconsistent image–text pairs multimodal medical logical deepfakes and study their detection as a defensive consistency-verification task. Image-only detectors cannot assess agreement with an unseen caption, while global visual self-attention incurs quadratic token interactions. We propose the Multimodal Bilateral Consistency Network (MM-BCNet), whose Text-Conditioned Vision Mamba (TC-VMamba) generates the selective state-space parameters Δ_i, $\mathbf{B}_i$, and $\mathbf{C}_i$ from each local visual token and a global text representation. This establishes direct text-to-parameter conditioning within the selective scan, enabling linear-complexity cross-modal recurrence across high-resolution visual sequences without token expansion. A single-query Semantic-Guided Affinity module supplies a spatial prior, and a Seeing Before Reasoning schedule learns visual forensic cues before multimodal fine-tuning. We construct a synthetic benchmark covering Semantic Misalignment and Visual Inpainting, with source-grouped splits and generation masks for edited-region localization, a task distinct from establishing semantic contradiction. Against baselines, MM-BCNet improves consistency classification, localization, and unseen-generator accuracy.

1 INTRODUCTION

Generative Artificial Intelligence (GenAI) is increasingly used in medical imaging research. Modern generative frameworks, including Latent Diffusion Models (LDMs) Rombach et al. (2022), can synthesize highly realistic biomedical images Wang et al. (2023). This capability creates a forensic risk when an image is locally manipulated or paired with altered text so that the two modalities communicate inconsistent medical information Zhang et al. (2024); Smeu et al. (2025); Li et al. (2025). As illustrated in Figure 1, we call these inconsistent image–text pairs *multimodal medical logical deepfakes*. They may arise through semantic modification of an otherwise valid description or through localized image generation that contradicts the associated text. Such inconsistencies can compromise the evidentiary value of biomedical publications and digital medical records. Our scope is defensive: we detect already-formed inconsistent pairs and localize supporting visual evidence; we do not propose a transferable attack or make clinical diagnoses.

Existing image-forgery detectors primarily model visual artifacts and therefore cannot decide whether a pristine image is semantically consistent with its accompanying text Wang et al. (2020); Li et al. (2026); Gye et al. (2025); Liu et al. (2024b). Multimodal Large Language Models (MLLMs) provide a natural route to image–text reasoning Shen et al. (2025); Wang et al. (2024); Liu et al. (2024a), but applying them to fine-grained medical forensics remains challenging. Global self-attention has quadratic complexity in the number of visual tokens, so practical systems often resize, pool, or tile high-resolution inputs; these operations can attenuate small manipulation traces Xu et al. (2023). Local-window attention reduces memory use but also limits direct interaction between distant windows. When visual evidence is weak or compressed, a multimodal model may rely disproportionately on linguistic priors and overlook a localized contradiction. These observations motivate a design that preserves fine-grained visual evidence while allowing the associated text to guide, rather than replace, visual analysis.

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

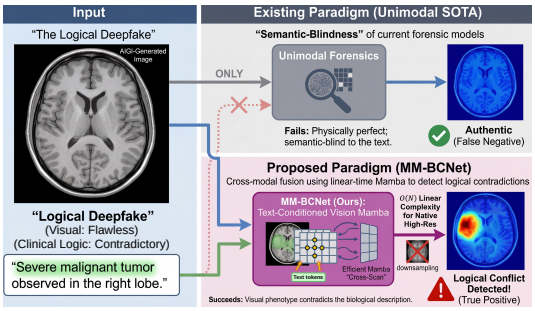

Figure 1: Detecting medical logical deepfakes. Top right: image-only detectors cannot evaluate whether a plausible image is consistent with its associated text. Bottom right: MM-BCNet combines text guidance with dense visual evidence to identify and localize image–text contradictions.

To address this problem, we propose the Multimodal Bilateral Consistency Network (MM-BCNet), organized around a *Seeing Before Reasoning* strategy. Its central component, Text-Conditioned Vision Mamba (TC-VMamba), uses a global text representation together with each local visual token to generate the selective SSM parameters Δ_i, $\mathbf{B}_i$, and $\mathbf{C}_i$ at every scan step. The design adds an explicit text-to-parameter path to an already input-dependent recurrence, rather than establishing that input-level text conditioning cannot affect selective memory. A Semantic-Guided Affinity (SGA) module first uses a single global text query to provide a spatial prior. TC-VMamba then scans the guided features and restores their patch-grid organization for decoding. Our efficiency motivation concerns the fusion path; the global-attention visual encoder remains part of the end-to-end cost.

To study this image–text consistency setting, we construct a benchmark using a two-stage curation strategy. Stage 1 uses BioFors to sensitize the visual encoder to physical manipulation cues. Stage 2 samples 50,000 biomedical image–text pairs from Open-PMC-18M and constructs negatives intended to introduce inconsistencies through two strategies: *Semantic Misalignment*, which modifies clinically relevant entities in the text, and *Visual Inpainting*, which inserts localized visual content while retaining the original text. These separate text-borne edits from image-borne manipulations; image editing alone is not a sufficient condition for semantic contradiction. Our main contributions are three-fold:

(1). We formulate multimodal medical logical-deepfake detection as a defensive image–text consistency task and construct a synthetic benchmark containing Semantic Misalignment and Visual Inpainting cases, with generation masks supervising edited-region localization.

(2). We propose MM-BCNet and TC-VMamba, which add explicit text-conditioned projections for the selective SSM parameters while retaining the visual patch-grid organization. The selective scan is linear in visual sequence length for fixed model width and state dimension.

(3). We introduce a lightweight single-query SGA module and a Seeing Before Reasoning training strategy that first learns visual forensic cues and then performs multimodal consistency reasoning.

2 Related Work

2.1 Unimodal AI-Generated Image Detection

Early forensic detectors target low-level physical signatures left by generative adversarial networks through spatial and frequency artifacts Wang et al. (2020). Subsequent methods address diffusion synthesis by quantifying trajectory reconstruction errors Wang et al. (2023), tracking color transition gradients Tan et al. (2023), or extracting generalized visual-linguistic representations Ojha et al. (2023). These unimodal detectors identify physical manipulation traces but cannot assess whether authentic visual evidence contradicts paired biomedical descriptions.

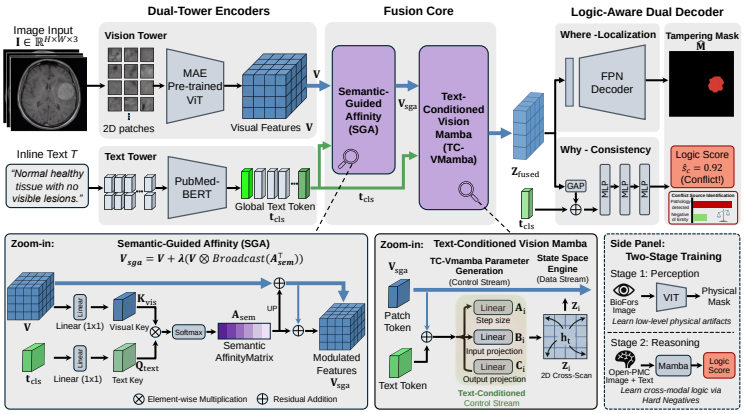

Figure 2: Overall architecture of MM-BCNet. Dual-tower features are fused via SGA and processed by TC-VMamba for cross-modal reasoning, followed by a Logic-Aware Dual Decoder predicting a tampering mask and a consistency score. Side panel: The Seeing Before Reasoning training paradigm.

2.2 Multimodal Forensics and Vision-Language Models

Multimodal forensic frameworks align visual features with textual representations to detect semantic misalignments between image content and associated claims Fan et al. (2024); Ji et al. (2026); Zhou et al. (2025). Applying vision-language architectures to biomedical forensics introduces computational and perceptual bottlenecks. The quadratic token complexity of global self-attention frequently forces input downsampling or aggressive pooling Xu et al. (2023), which attenuates localized editing traces. Incomplete visual grounding encourages undue reliance on language priors Chen et al. (2025), underscoring the need for dense cross-modal interaction without quadratic expansion.

2.3 State Space Models in Medical Imaging

State Space Models establish linear-complexity sequence modeling for long-range visual representations Gu & Dao (2023); Liu et al. (2024c); Pei et al. (2025). Medical architectures exploit selective scanning for dense anatomical prediction across 2D and 3D modalities Ma et al. (2024b); Xing et al. (2025), while BioTamperNet Nandi & Natarajan (2026) targets physical manipulation detection via visual affinity cues. In cross-modal domains, ReMamber Yang et al. (2024) combines visual and textual sequences through bidirectional channel and spatial twisting. In contrast to sequence-twisting pipelines or indirect input modulation, MM-BCNet combines a single-query spatial affinity prior with explicit text conditioning injected directly into the selective recurrent parameters Δ_i, $\mathbf{B}_i$, and $\mathbf{C}_i$.

3 Method

To tackle logical deepfakes in biomedical multimedia, we propose the Multimodal Bilateral Consistency Network (**MM-BCNet**). Unlike image-only forensic models, MM-BCNet evaluates whether a biomedical image is consistent with its associated textual description while retaining a visual localization output for image-borne manipulations.

3.1 Architecture Overview

Formally, given a biomedical image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and its associated text T, MM-BCNet learns a mapping $\mathcal{F}(\mathbf{I}, T) \rightarrow (\hat{\mathbf{M}}, \hat{s}_c)$ with two complementary outputs. The soft mask $\hat{\mathbf{M}} \in [0,1]^{H\times W}$ scores image manipulation at each pixel and can be thresholded for edited-region localization. The bounded score $\hat{s}_c \in [0,1]$ is trained to predict the benchmark's inconsistency label; it is not established as a calibrated probability of a medically adjudicated contradiction.

Task boundary. Image provenance and semantic consistency are distinct: an edited image can agree with its caption, while an unedited image can contradict a modified caption. The benchmark treats source correspondences as consistent and assigns intended inconsistency labels to constructed variants. These construction labels operationalize the task but do not independently establish the semantic validity of every pair. Generation masks identify edited pixels, not necessarily regions that contradict the text.

As illustrated in Figure 2, an MAE-initialized ViT-B/16 visual tower maps $\mathbf{I}$ to patch features $\mathbf{V} \in \mathbb{R}^{L\times d}$, where $L = (H/16)(W/16)$ and d is the embedding dimension. Concurrently, PubMedBERT maps T to token features $\mathbf{T} \in \mathbb{R}^{K\times d}$ and a global representation $\mathbf{t}_{cls}$. SGA first produces text-guided visual features $\mathbf{V}_{sga}$, after which TC-VMamba performs text-conditioned selective scanning. A logic-aware dual decoder then predicts $\hat{\mathbf{M}}$ and $\hat{s}_c$. The linear-complexity statement in this work refers specifically to the TC-VMamba selective scan with respect to L; end-to-end computation also includes the visual and text encoders and is therefore evaluated empirically.

3.2 Semantic-Guided Affinity (SGA) Module

Biomedical images often contain dense and repetitive textures, such as cell clusters and electrophoresis bands. SGA uses the global text representation as a soft spatial prior so that text-relevant visual locations receive greater weight before state-space modeling, while the residual path preserves the original visual features.

We project the global text feature as a single query and the visual features as spatial keys:

$$\mathbf{Q}_{text} = \mathbf{t}_{cls}\mathbf{W}_q, \qquad \mathbf{K}_{vis} = \mathbf{V}\mathbf{W}_k, \tag{1}$$

where $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d\times d}$ are learnable projections. The resulting semantic affinity vector $\mathbf{A}_{sem} \in \mathbb{R}^{1\times L}$ assigns a normalized relevance weight to each visual location:

$$\mathbf{A}_{sem} = \text{Softmax}\left(\frac{\mathbf{Q}_{text}\mathbf{K}_{vis}^{\top}}{\sqrt{d}}\right) \tag{2}$$

Rather than replacing the visual features, we use residual multiplicative modulation:

$$\mathbf{V}_{sga} = \mathbf{V} \odot \left(\mathbf{1} + \lambda \cdot \text{Broadcast}(\mathbf{A}_{sem}^{\top})\right), \tag{3}$$

where $\odot$ is the Hadamard product and the learnable scale λ is initialized to zero. This starts SGA as an identity mapping and lets the model learn the strength of text guidance. After projection, the single-query affinity product costs $O(Ld)$ and stores an $O(L)$ vector; the spatial key projection adds $O(Ld^2)$. Cross-attention over K text tokens instead uses an $O(KLd)$ interaction product and an $O(KL)$ map, in addition to its projections. For fixed K, that interaction is also linear in L. SGA therefore reduces the interaction-map size by pooling text, rather than changing the asymptotic order relative to fixed-length cross-attention.

3.3 Text-Conditioned Vision Mamba

Global visual self-attention has quadratic token interactions, whereas selective scanning is linear in visual sequence length for fixed model dimensions. TC-VMamba keeps visual tokens in their scan order and supplies the global text representation directly to the selective parameter projections, without inserting text tokens into the visual grid.

SSM Preliminaries. For clarity, we present the recurrence for one feature channel; the vector implementation applies the selective SSM in parallel across the d feature channels. A continuous-time SSM maps $x(t) \in \mathbb{R}$ to $y(t) \in \mathbb{R}$ through $\mathbf{h}(t) \in \mathbb{R}^N$:

$$\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}x(t), \qquad y(t) = \mathbf{C}^{\top}\mathbf{h}(t), \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^N$. With a zero-order hold and input-dependent parameters at position i, the discrete update is

$$\bar{\mathbf{A}}_i = \exp(\Delta_i \mathbf{A}), \qquad \bar{\mathbf{B}}_i = \left(\int_0^{\Delta_i} \exp(\tau\mathbf{A})\, d\tau\right)\mathbf{B}_i, \tag{5}$$

$$\mathbf{h}_i = \bar{\mathbf{A}}_i \mathbf{h}_{i-1} + \bar{\mathbf{B}}_i x_i, \qquad y_i = \mathbf{C}_i^\top \mathbf{h}_i. \tag{6}$$

Thus, $\Delta_i \in \mathbb{R}_+$ controls the discretized transition, while $\mathbf{B}_i, \mathbf{C}_i \in \mathbb{R}^N$ control the position-dependent state input and output.

Text-Conditioned Recurrent Dynamics. Let $\mathbf{v}_i \in \mathbb{R}^d$ denote the i-th visual token of $\mathbf{V}_{sga}$ and $\mathbf{t}_{cls} \in \mathbb{R}^d$ represent the projected global text representation. TC-VMamba derives the discretization parameter Δ_i alongside the state matrices $\mathbf{B}_i$ and $\mathbf{C}_i$ through joint affine transformations:

$$\Delta_i = \text{Softplus}\big(\mathbf{W}_{\Delta,v}\mathbf{v}_i + \mathbf{W}_{\Delta,t}\mathbf{t}_{cls} + b_\Delta\big), \tag{7}$$

$$\mathbf{B}_i = \mathbf{W}_{B,v}\mathbf{v}_i + \mathbf{W}_{B,t}\mathbf{t}_{cls} + \mathbf{b}_B, \qquad \mathbf{C}_i = \mathbf{W}_{C,v}\mathbf{v}_i + \mathbf{W}_{C,t}\mathbf{t}_{cls} + \mathbf{b}_C. \tag{8}$$

The underlying continuous system maps to discrete recurrence via zero-order hold:

$$\bar{\mathbf{A}}_i = \exp(\Delta_i \mathbf{A}), \qquad \bar{\mathbf{B}}_i = \left(\int_0^{\Delta_i} \exp(\tau \mathbf{A})\, d\tau \right) \mathbf{B}_i \approx \Delta_i \mathbf{B}_i. \tag{9}$$

Because the diagonal state transition matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ has strictly negative entries governing decay, the text embedding $\mathbf{t}_{cls}$ modulates temporal memory retention non-linearly through Δ_i. High contextual relevance shifts Δ_i upward, forcing $\bar{\mathbf{A}}_i \to 0$ to purge irrelevant visual history, while lower values preserve long-range anatomical representations across scan lines. Simultaneously, the state input term $\bar{\mathbf{B}}_i x_i \approx \Delta_i(\mathbf{W}_{B,v}\mathbf{v}_i + \mathbf{W}_{B,t}\mathbf{t}_{cls} + \mathbf{b}_B)x_i$ induces multiplicative interactions among the visual token x_i, spatial context $\mathbf{v}_i$, and linguistic guidance $\mathbf{t}_{cls}$. This parameterization provides independent text-controlled projections for memory decay, state admission, and state readout.

To preserve the 2D token layout, we implement a 2D cross-modal scan over four directions: top-left to bottom-right, the reverse direction, top-right to bottom-left, and its reverse. The four direction-specific outputs are returned to raster order and summed to form $Z_{1D} \in \mathbb{R}^{L\times d}$. We then reshape Z_{1D} into $Z_{fused} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times d}$ so that the downstream decoder operates on the original patch-grid organization.

Complexity scope. For fixed width d, state dimension N, and number of blocks, parameter projection and selective scanning scale linearly with L; four scan directions change only a constant factor. This does not remove the $O(L^2 d)$ attention term in the ViT visual encoder or establish a latency or memory advantage over an optimized cross-attention implementation.

3.4 Logic-Aware Dual Decoder

The dual-head decoder transforms the spatially-restored multimodal feature map Z_{fused} into discriminative forensic decisions.

For the **Tampering Localization Head**, we employ a lightweight FPN-style decoder that progressively refines and upsamples Z_{fused} to the original $H \times W$ resolution. A 1×1 convolution followed by a Sigmoid activation yields the soft pixel-level mask:

$$\hat{\mathbf{M}} = \sigma\big(\text{FPN}(Z_{fused})\big) \tag{10}$$

For the **Consistency Reasoning Head**, we perform global average pooling on Z_{fused}, concatenate the pooled feature with $\mathbf{t}_{cls}$, and pass it through a Multi-Layer Perceptron (MLP) to compute a bounded inconsistency score:

$$\hat{s}_c = \sigma\big(\text{MLP}(\text{GlobalAvgPool}(Z_{fused}) \oplus \mathbf{t}_{cls})\big) \tag{11}$$

3.5 Seeing Before Reasoning: Two-Stage

We use a *Seeing Before Reasoning* two-stage strategy to learn visual forensic features before multimodal fine-tuning. This schedule is intended to reduce premature reliance on text cues; the training order alone does not guarantee the absence of shortcuts.

Stage 1: Perception Pre-training. We mask the text by setting its representation to zero and train the visual encoder, the image-only pathway through TC-VMamba, and the localization decoder on

Table 1: Quantitative evaluation on the Logical Deepfakes Benchmark across 20,000 test instances. Overall metrics reflect micro-evaluations across the pooled test collection. Baseline models in Group 2 are fine-tuned on the Stage-2 training split using identical LoRA rank configurations. * indicates medical-domain initialization.

Method	Modality	Semantic Misalignment (Real Img + Fake Txt)			Visual Inpainting (Fake Img + Real Txt)			Overall Benchmark		
		ACC (%)	AUC (%)	mAP (%)	ACC (%)	AUC (%)	mAP (%)	ACC (%)	AUC (%)	mAP (%)
Group 1: Unimodal Image Forensics (Text-Blind)										
CNNSpot Wang et al. (2020)	Image	48.23	50.11	51.04	54.12	56.34	55.89	51.18	53.23	53.47
FreDect Frank et al. (2020)	Image	49.15	49.88	50.56	56.78	58.91	57.22	52.97	54.40	53.89
GramNet Liu et al. (2020)	Image	50.04	51.23	51.45	58.45	60.12	59.34	54.25	55.68	55.40
Fusing Ju et al. (2022)	Image	51.34	52.45	51.89	61.23	63.45	62.78	56.29	57.95	57.34
LNP Liu et al. (2022a)	Image	49.88	50.34	50.67	60.56	62.11	61.45	55.22	56.23	56.06
LGrad Tan et al. (2023)	Image	50.45	51.12	51.56	65.34	67.89	66.12	57.90	59.51	58.84
UnivFD Ojha et al. (2023)	Image	51.67	52.88	52.14	68.12	70.45	69.34	59.90	61.67	60.74
NPR Tan et al. (2024)	Image	49.21	50.56	50.88	69.45	71.23	70.56	59.33	60.90	60.72
FatFormer Liu et al. (2024b)	Image	50.89	51.45	51.67	70.11	72.56	71.89	60.50	62.01	61.78
RealNet Li et al. (2026)	Image	51.12	52.01	51.90	71.34	73.67	72.45	61.23	62.84	62.18
SFLD Gye et al. (2025)	Image	52.34	53.67	52.89	74.56	76.89	75.12	63.45	65.28	64.01
Group 2: Multimodal Large Language Models (Target Fine-Tuned)										
CLIP (ViT-L/14) Zhao et al. (2025)	Img+Txt	73.45	75.12	74.56	58.67	60.23	59.12	66.06	67.68	66.84
BLIP-2 Li et al. (2023b)	Img+Txt	76.12	78.34	77.89	62.34	64.56	63.89	69.23	71.45	70.89
Qwen2-VL Wang et al. (2024)	Img+Txt	80.56	82.45	81.12	65.78	67.12	66.34	73.17	74.79	73.73
LLaVA-1.5 Liu et al. (2024a)	Img+Txt	82.34	84.56	83.45	68.45	70.34	69.12	75.40	77.45	76.29
LLaVA-Med* Li et al. (2023a)	Img+Txt	86.78	88.12	87.56	72.11	74.56	73.89	79.45	81.34	80.73
MM-BCNet (Ours)	**Img+Txt**	**89.24**	**91.15**	**90.45**	**84.76**	**86.32**	**85.88**	**87.00**	**88.74**	**88.17**

BioFors. This stage learns visual anomaly cues for physical manipulations such as copy-move, splicing, and local enhancement:

$$\mathcal{L}_{stage1} = \mathcal{L}_{BCE}(\hat{\mathbf{M}}, \mathbf{M}_{gt}) + \gamma \mathcal{L}_{recon} \tag{12}$$

where $\mathcal{L}_{recon}$ is a self-reconstruction objective that encourages preservation of fine visual detail.

Stage 2: Logic-Aware Contrastive Fine-tuning. We freeze the Stage-1 visual encoder and fine-tune SGA, TC-VMamba, the consistency head, and the localization decoder on the Open-PMC-18M-derived benchmark. We construct hard negatives through two strategies: (1) *Semantic Misalignment*, where a source image is paired with text containing a clinically meaningful entity change (e.g., left lobe to right lobe); and (2) *Visual Inpainting*, where localized image content is generated while the source text is retained. The Stage 2 optimization objective balances consistency verification, cross-modal alignment, and spatial localization:

$$\mathcal{L}_{stage2} = \mathcal{L}_{BCE}(\hat{s}_c, y_{logic}) + \beta \mathcal{L}_{InfoNCE}(\mathbf{Z}_{global}, \mathbf{t}_{cls}) + \mu \mathcal{L}_{mask}, \tag{13}$$

where $y_{logic} \in \{0, 1\}$ designates pair consistency. For the B_a authentic pairs in a batch, normalized representations $\mathbf{z}_i = \text{norm}(g_z(\mathbf{Z}^i_{global}))$ and $\mathbf{u}_i = \text{norm}(g_t(\mathbf{t}^i_{cls}))$ are projected through learned heads g_z and g_t. The contrastive term is defined as:

$$\mathcal{L}_{InfoNCE} = -\frac{1}{B_a} \sum_{i=1}^{B_a} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{u}_i / \tau)}{\sum_{j=1}^{B_a} \exp(\mathbf{z}_i^\top \mathbf{u}_j / \tau)}. \tag{14}$$

Here, $\mathbf{Z}_{global} = \text{GlobalAvgPool}(Z_{fused})$ aggregates the text-conditioned visual tokens into a global multimodal descriptor. Aligning $\mathbf{Z}_{global}$ with $\mathbf{t}_{cls}$ enforces semantic coherence within the fused cross-modal space. Although Z_{fused} has already interacted with $\mathbf{t}_{cls}$ inside TC-VMamba, the model is prevented from learning a trivial text-routing shortcut because the simultaneous consistency and mask objectives strictly require faithful visual grounding. We formally validate this visual dependency via input-ablation and permutation stress tests in the Supplementary Material.

4 Experiments

4.1 Experimental Setup

We use a two-stage data protocol that separates visual-forensic pre-training from multimodal consistency fine-tuning.

Stage 1: Perception Pre-training Dataset. We use BioFors Sabir et al. (2021), which contains more than 47,000 biomedical images from microscopy, blot/gel, and macroscopic domains, with physical

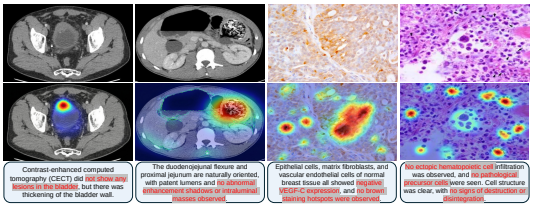

Figure 3: Qualitative evidence across CT, immunohistochemistry, and H&E pathology examples. The paired text negates selected findings (red), while Grad-CAM highlights visual regions contributing to the inconsistency prediction.

manipulations such as copy-move, splicing, and contrast enhancement. We follow its standard partitions, using only the training partition to optimize the visual encoder and localization decoder; the text input is masked during this stage.

Stage 2: Logical Deepfake Evaluation Benchmark. We sample 50,000 source image–text pairs from Open-PMC-18M Baghbanzadeh et al. (2025) and construct contradictory pairs using two strategies. (1) *Semantic Misalignment* (Real Img + Fake Txt): GPT-4 modifies a clinically relevant entity in the source text while preserving its surrounding syntax and fluency. (2) *Visual Inpainting* (Fake Img + Real Txt): given a region prompt from the generation pipeline, MedSAM Ma et al. (2024a) produces an anatomically localized mask. A latent diffusion model fine-tuned on medical imagery then generates content within that mask while the source text remains unchanged; the same mask is retained as pixel-level ground truth. A CLIP-based quality filter removes low-similarity or visibly degraded generations.

4.1.1 Implementation Details

MM-BCNet is implemented in PyTorch. The text tower is initialized from PubMedBERT Gu et al. (2021), and the visual tower uses an MAE-initialized ViT-B/16 encoder followed by SGA and TC-VMamba. TC-VMamba uses an SSM state dimension of $N = 16$ and a spatial scan stride of one token. The loss weights are $\gamma = 0.3$ for reconstruction in Stage 1 and $\beta = 0.5$ for InfoNCE and $\mu = 1.0$ for the Visual-Inpainting mask term in Stage 2; the SGA scale λ is initialized to zero and learned end-to-end. We use a batch size of 32 and AdamW with weight decay 10^{-4}. Stage 1 trains the visual encoder, the image-only TC-VMamba pathway, and the localization decoder for 50 epochs with learning rate 1×10^{-4}. Stage 2 freezes the Stage-1 visual encoder and fine-tunes SGA, TC-VMamba, the consistency head, and the localization decoder for 30 epochs with learning rate 5×10^{-5}. Both stages use five warmup epochs followed by cosine-annealed learning rates. Experiments are conducted on a server with four NVIDIA A100 80GB GPUs.

Table 2: Cross-generator generalization performance (ACC %). Models are trained exclusively on SD v1.4 and evaluated zero-shot on unseen generators. Best and second-best results are **bold** and underlined.

Method	Modality	Intra-Domain	Cross-Domain (Unseen Generators)				Unseen Avg. (↑)
		SD v1.4 (Train)	StyleGAN3	DALL-E 3	Midjourney v6	Med-GAN	
Group 1: Unimodal Image Forensics (Artifact-Dependent)							
LGrad Tan et al. (2023)	Image	81.34	58.45	55.12	52.34	56.89	55.70
UnivFD Ojha et al. (2023)	Image	86.12	64.23	61.45	54.67	60.12	60.12
NPR Tan et al. (2024)	Image	88.56	59.34	57.89	51.22	53.45	55.48
SFLD Gye et al. (2025)	Image	89.45	66.89	62.34	57.45	61.23	61.98
Group 2: Multimodal Large Language Models (Semantic-Dependent)							
CLIP (ViT-L/14) Zhao et al. (2025)	Img+Txt	70.12	68.34	67.45	65.12	66.89	66.95
Qwen2-VL Wang et al. (2024)	Img+Txt	76.45	74.12	73.56	70.89	71.34	72.48
LLaVA-Med Li et al. (2023a)	Img+Txt	83.12	81.45	79.56	76.23	78.89	79.03
MM-BCNet (Ours)	**Img+Txt**	**91.34**	**89.12**	**87.45**	**85.67**	**86.34**	**87.15**

4.1.2 Evaluation Metrics

We evaluate the performance across two dimensions. For Image-level Consistency Verification (whether the pair is a logical deepfake), we report the standard classification metrics: Accuracy

(ACC), Area Under the Receiver Operating Characteristic Curve (AUC), and Mean Average Precision (mAP). For Pixel-level Tampering Localization (where the manipulation occurred), we evaluate the generated mask $\hat{\mathbf{M}}$ using the pixel-level F1-score and Intersection over Union (IoU) against the ground-truth perturbation masks.

4.2 Main Results on Logical Deepfake Detection

We evaluate MM-BCNet against 16 baselines across two logical deepfake scenarios, namely Semantic Misalignment and Visual Inpainting, as detailed in Table 1. Unimodal image detectors such as UnivFD Ojha et al. (2023) and SFLD Gye et al. (2025) exhibit near-chance performance on Semantic Misalignment because physical pixel grids remain authentic. Multimodal baselines fine-tuned on target distributions show improved sensitivity to text alterations but degrade when identifying localized image edits; for example, LLaVA-Med Li et al. (2023a) achieves 72.11% accuracy on Visual Inpainting. MM-BCNet obtains 89.24% on Semantic Misalignment and 84.76% on Visual Inpainting, reaching 87.00% accuracy across the 20,000 pooled test pairs. Preserving uncompressed spatial token structures through TC-VMamba enables joint tracking of local manipulation boundaries and cross-modal semantic discrepancies.

As a small external check beyond synthetic construction, we additionally evaluate a frozen MM-BCNet (threshold fixed on the Stage-2 validation split) on Integrity-100, an expert-adjudicated set of 100 open-access publication figures with contested image–text consistency (supplementary material). Accuracy drops relative to the synthetic benchmark, consistent with the failure taxonomy, and we treat the set as a transfer probe rather than a deployment study.

Qualitative inspections corroborate this tracking behavior by revealing the spatial focus of the consistency reasoning head. Figure 3 demonstrates spatial localization of contradictions across radiological and histological modalities. When captions falsely assert the absence of pathology, Grad-CAM heatmaps highlight the focal bladder lesion and intraluminal gastric mass on CT. Histological evaluations mirror this selectivity, isolating brown VEGF-C staining hotspots and infiltrated precursor cells to ground consistency decisions.

4.3 Cross-Generator Generalization

To assess sensitivity to generator-specific artifacts, we restrict this experiment to Visual Inpainting, where generator identity is a controlled variable. All models are trained on the Stable Diffusion v1.4 (SD v1.4) subset only and tested without adaptation on StyleGAN3, DALL-E 3, Midjourney v6, and Med-GAN. Because these generators do not all expose a mask-conditioned medical inpainting interface, each synthesizes a full image from the source caption and the region defined by the same MedSAM mask is composited into the authentic scan, so only generator identity varies. Each unseen-generator set is balanced (1,500 inconsistent and 1,500 authentic pairs), and the threshold is fixed on the SD v1.4 validation split.

Table 2 shows a larger drop for the unimodal baselines, whose unseen-generator averages range from 55.48% to 61.98%. LLaVA-Med reaches 79.03%, while MM-BCNet obtains the best reported average of 87.15%. Intra-domain accuracies here exceed the Visual-Inpainting column of Table 1 because this suite fixes one generator and one blending procedure, so the two settings are not directly comparable. The result is consistent with cross-modal consistency providing a cue less tied to a particular generator, although evaluation on additional generators and independently collected data remains necessary.

4.4 Pixel-Level Tampering Localization

To fulfill our evaluation protocol, we assess pixel-level forgery localization exclusively on the Visual Inpainting subset, whose generation masks provide ground truth. All methods, including the baselines, receive the same mask supervision on the same Stage-2 split and are evaluated on the same held-out source groups, so the comparison reflects architecture rather than supervision. Table 3 compares MM-BCNet against nine state-of-the-art baselines using pixel-level F1-score and IoU.

Table 3 shows that MM-BCNet improves over the strongest listed baseline from 61.45% to 78.62% F1 and from 48.33% to 65.84% IoU. Unimodal methods can exploit physical traces but cannot use

Table 3: Pixel-level tampering localization performance on the Visual Inpainting scenario. Unimodal baselines strictly rely on visual artifacts, while Multimodal baselines incorporate text guidance. Best results are in **bold**, and second-best are underlined.

Method	Modality	F1-score (%) ↑	IoU (%) ↑
Group 1: Unimodal Image Forensics (Artifact-Dependent)			
SPAN Hu et al. (2020)	Image Only	32.14	21.56
ManTraNet Wu et al. (2019)	Image Only	34.65	23.88
OSN Wu et al. (2022)	Image Only	38.41	27.12
CAT-Net Kwon et al. (2021)	Image Only	41.22	29.34
PSCC-Net Liu et al. (2022b)	Image Only	42.87	31.05
MVSS-Net Chen et al. (2021)	Image Only	44.50	33.18
HiFi-Net Guo et al. (2023)	Image Only	51.34	39.82
DiffForensics Yu et al. (2024)	Image Only	54.12	42.67
Group 2: Multimodal Forensics (Semantic-Dependent)			
FakeShield Xu et al. (2024)	Img + Txt	61.45	48.33
MM-BCNet (Ours)	**Img + Txt**	**78.62**	**65.84**

the paired description to disambiguate visually plausible tissue, whereas SGA and TC-VMamba add complementary semantic and spatial evidence under identical supervision. Because the ground truth is the generation mask, this is forensic localization of image-borne manipulations, not clinical lesion segmentation.

4.5 Ablation Studies

Table 4: Ablation studies on architectural components and training stages. Evaluated across the complete benchmark. PT denotes Stage-1 perception pre-training.

Model Variant	SGA	PT	ACC (%)	AUC (%)
Baseline Architecture (Standard Cross-Attention)				
(a) ViT-B/16 (Image-Only)	✗	✗	63.45	65.28
(b) ViT-B/16 + Cross-Attn	✗	✗	78.12	80.45
(c) ViT-B/16 + Cross-Attn	✓	✗	81.34	83.67
(d) ViT-B/16 + Cross-Attn	✓	✓	84.23	85.89
Proposed Architecture (TC-VMamba)				
(e) TC-VMamba (Image-Only)	✗	✗	65.78	67.34
(f) TC-VMamba (Concat Fusion)	✗	✗	82.56	84.12
(g) TC-VMamba (SGA Fusion)	✓	✗	85.12	86.89
(h) MM-BCNet (Full Model)	✓	✓	**87.00**	**88.74**

To validate the contribution of our architectural innovations and the two-stage training paradigm, we conduct ablation studies on the complete Logical Deepfakes Benchmark (Table 4). Comparing (f) and (g), SGA improves accuracy by 2.56 points over concatenation, supporting the value of a text-derived spatial prior. Under the matched settings of (d) and (h), replacing cross-attention fusion with TC-VMamba improves accuracy by 2.77 points. Removing Stage-1 perception pre-training reduces accuracy from 87.00% to 85.12% (h versus g). These controlled variants isolate the proposed fusion and training components; they do not imply that the end-to-end architecture is linear in image resolution.

5 Conclusion

MM-BCNet provides a defensive framework for detecting multimodal medical logical deepfakes by coupling the semantic spatial prior of SGA with text-conditioned selective recurrence in TC-VMamba. Direct parameter generation for Δ_i, $\mathbf{B}_i$, and $\mathbf{C}_i$ yields substantial gains over unimodal and vision-language baselines across consistency classification, pixel-level tampering localization, and unseen-generator generalization. By reformulating cross-modal fusion as text-conditioned state transitions, MM-BCNet demonstrates that selective state-space models can effectively bridge low-level forensic evidence and high-level logical reasoning. We envision this paradigm serving as a robust defensive foundation for safeguarding biomedical multi-modal integrity against increasingly capable generative models.

REPRODUCIBILITY STATEMENT

The Supplementary Material details article-level PubMed Central partitions, pixel-level manipulation masks, and complete training hyperparameters to support exact experimental replication. Operator-level computational profiling, reporting peak memory allocations and execution latencies on NVIDIA GPUs, documents resource efficiency. Diagnostic stress tests involving null-image ablations, cross-sample feature shuffling, and four compositing pipelines rule out linguistic shortcuts, with all primary metrics validated across three random seeds.

AI USE STATEMENT

We used GPT-4 to generate semantic-misalignment captions and a latent diffusion model to construct Visual-Inpainting samples for the benchmark. All generated content was automatically filtered and manually reviewed. We did not use generative AI to produce the manuscript, figures, code, or experimental analysis. The authors reviewed all AI-assisted outputs and take full responsibility for the final content and results.

REFERENCES

Negin Baghbanzadeh, Sajad Ashkezari, Elham Dolatabadi, and Arash Afkanpour. Open-pmc-18m: A high-fidelity large scale medical dataset for multimodal representation learning. *arXiv preprint arXiv:2506.02738*, 2025.

Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, and Zheng Feng. Enhancing uncertainty modeling with semantic graph for hallucination detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23586–23594, 2025.

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*, pp. 14185–14193, 2021.

Yuming Fan, Dongming Yang, Jiguang Zhang, Bang Yang, and Yuexian Zou. Fake-gpt: Detecting fake image via large language model. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 122–136. Springer, 2024.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3155–3165, 2023.

Seoyeon Gye, Junwon Ko, Hyounguk Shon, Minchan Kwon, and Junmo Kim. Reducing the content bias for ai-generated image detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 399–408. IEEE, 2025.

Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *European conference on computer vision*, pp. 312–328. Springer, 2020.

Yikun Ji, Yan Hong, Qi Fan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, Jianfu Zhang, et al. Fakexplain: Ai-generated image detection via human-aligned grounded reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3465–3469. IEEE, 2022.

Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 375–384, 2021.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 626–637. Springer, 2025.

Shuaibo Li, Laixin Zhang, Wei Ma, Jianwei Guo, Shibiao Xu, Zhijie Qiu, and Hongbin Zha. Realnet: Efficient and unsupervised detection of ai-generated images via real-only representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38889–38898, 2026.

Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European conference on computer vision*, pp. 95–110. Springer, 2022a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024b.

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, 2022b.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024c.

Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8060–8069, 2020.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024a.

Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024b.

Soumyaroop Nandi and Prem Natarajan. Biotampernet: Affinity-guided state-space model detecting tampered biomedical images. In *International Conference on Learning Representations (ICLR)*, 2026.

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*, pp. 24480–24489, 2023.

Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 6443–6451, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 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*, pp. 10963–10973, 2021.

Chen Shen, Chunfeng Lian, Wanqing Zhang, Fan Wang, Jianhua Zhang, Shuanliang Fan, Xin Wei, Gongji Wang, Kehan Li, Hongshu Mu, et al. Large-vocabulary forensic pathological analyses via prototypical cross-modal contrastive learning. *Nature Communications*, 16(1):6773, 2025.

Stefan Smeu, Elisabeta Oneata, and Dan Oneata. Declip: Decoding clip representations for deepfake localization. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 149–159, February 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

Chuangchuang Tan, 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*, pp. 28130–28139, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

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, October 2023.

Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13440–13449, 2022.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9543–9552, 2019.

Zhaohu Xing, Tian Ye, Yijun Yang, Du Cai, Baowen Gai, Xiao-Jian Wu, Feng Gao, and Lei Zhu. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Hongming Xu, Qi Xu, Fengyu Cong, Jeonghyun Kang, Chu Han, Zaiyi Liu, Anant Madabhushi, and Cheng Lu. Vision transformers for computational histopathology. *IEEE Reviews in Biomedical Engineering*, 17:63–79, 2023.

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. *arXiv preprint arXiv:2410.02761*, 2024.

Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. ReMamber: Referring image segmentation with mamba twister. In *European Conference on Computer Vision (ECCV)*, 2024.

Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12765–12774, 2024.

Guisheng Zhang, Mingliang Gao, Qilei Li, Wenzhe Zhai, and Gwanggil Jeon. Multi-modal generative deepfake detection via visual-language pretraining with gate fusion for cognitive computation. *Cognitive Computation*, 16(6):2953–2966, 2024.

Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A survey. *Medical Image Analysis*, 102:103551, 2025.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025.