Dual-Comb Ghost Imaging with Transformer-Based Reconstruction for Optical Fiber Endomicroscopy

David Dang[1,3] **Myoung-Gyun Suh**[2] **Maodong Gao**[2] **Byoung Jun Park**[2]
Beyonce Hu[1] **Yucheng Jin**[1] **Wilton J.M. Kort-Kamp**[3] **Ho Wai (Howard) Lee**[1]
[1]University of California, Irvine [2]NTT Physics and Informatics Laboratories
[3]Los Alamos National Laboratory
{dangd5,beyonceh,yuchej9,Howardhw.lee}@uci.edu
{myoung-gyun.suh,byoungjun.park}@ntt-research.com, mgao@caltech.edu
kortkamp@lanl.gov

Abstract

Endoscopic imaging is indispensable for visualizing internal organs, yet conventional systems remain bulky and costly because they rely on large, multi-element optics, which limits their ability to access and image certain areas of the body. Achieving high-quality endomicroscopy with hundred micron-scale and inexpensive hardware remains a grand challenge. Optical fibers offer a sub-millimeter-scale imaging conduit that could meet this need, but existing fiber-based approaches typically require either raster scanning or multicore bundles, which limit resolution and speed of imaging. In this work, we overcome these limitations by combining dual-comb interferometry with optical ghost imaging and advanced algorithm. Optical frequency combs enable precise and parallel speckle illumination via wavelength-division multiplexing through a single-core fiber, while our dual-comb compressive ghost imaging approach enables snapshot detection of bucket-sum signals using a single-pixel detector, eliminating the need for both spatial and spectral scanning. To reconstruct images from these highly compressed measurements, we introduce **Optical Ghost-GPT**, a transformer-based image reconstruction model that enables fast, high-fidelity recovery at low sampling ratios. Our dual-comb ghost imaging approach, combined with the novel algorithm, outperforms classical ghost imaging techniques in both speed and accuracy, enabling real-time, high-resolution endoscopic imaging with a significantly reduced device footprint. This advancement paves the way for non-invasive, high-resolution, low-cost endomicroscopy and other sensing applications constrained by hardware size and complexity.

1 Introduction

Endoscopes are an important tool in modern medical diagnostics, enabling direct visualization of internal organs and tissues. From gastrointestinal investigations to bronchoscopies and laparoscopies, endoscopic techniques are widely used for both diagnostic and therapeutic purposes. Traditional endoscopes typically consist of a long, flexible or rigid tube equipped with a light source and a camera system to capture real-time images [1]. Despite their essential role in clinical practice, conventional endoscopes face several critical challenges: **(a)** The complex electronics and imaging systems increases an endoscope's size to the order of millimeters to centimeters— resulting in invasive and uncomfortable procedures on a patient; **(b)** Moreover, the imaging capabilities are typically limited by the size of the optics and cameras and prevent the imaging of small body parts; **(c)** Additionally, these imaging systems play a role in the large cost of endoscopes, with average prices in the range of several tens of thousands of dollars [2].

Fiber endomicroscopy offers a markedly smaller alternative. By exploiting single-core, multicore, or multimode optical fibers with outer diameters below 500 micron, it is possible to deliver light and collect signals through a conduit scarcely thicker than a human hair. Early demonstrations, including confocal[3] and multiphoton fluorescence microendoscopes[4], established that cellular-scale imaging could be achieved with a sub-millimeter footprint. Subsequent advances such as scanning-fiber endoscopes and microelectromechanical (MEMS) scanners improved field of view[5] , while computational approaches have enabled lensless, holographic, and light-field reconstructions through flexible probes as thin as 200 micron[6, 7]. Despite these successes, most fiber-based systems still require either mechanical components such as MEMS scanner or rotating torque coil for raster scanning or coherent fiber bundles[8] whose inter-core spacing limits resolution, operation speed, and reduces fill factor.

Ghost imaging provides a way to bypass these bottlenecks by replacing pixelated cameras with correlated intensity measurements from a single-pixel (bucket) detector[9]. In a typical implementation, a sequence of known illumination patterns are projected onto the sample, and statistical correlations between these patterns and the measured total intensities are used to computationally reconstruct the image. Ghost imaging has demonstrated advantages in low light conditions due to improved signal-to-noise ratios[10, 11, 12], the ability to image through scattering media[13, 14], and compatibility with single-pixel detectors - making it well suited for fiber-based imaging, where 2D detector arrays are impractical[15, 16]. However, classical ghost imaging remains slow because each pattern must be projected sequentially, and iterative reconstruction algorithms often converge slowly, are prone to low image fidelity, or stall at low sampling ratios.

In this work, we combine dual optical frequency comb (dual-comb, for short) interferometry[17, 18, 19] with compressive ghost imaging to realize snapshot speckle imaging through a single-core fiber, eliminating the need for slow spatial or spectral scanning while preserving a minimal footprint. For image recovery, we introduce **Optical Ghost-GPT**, a transformer-based reconstruction model that enables real-time, high-fidelity imaging. Our contributions include: **(1)** First demonstration of optical fiber-based ghost imaging using a hardware-software co-design that combines dual-comb interferometry and deep learning, **(2)** superior reconstruction speed and resolution through the hardware-software co-design approach, **(3)** a robust transformer-based framework that maintains performance in noisy environments.

2 Background on Single-Pixel and Ghost Imaging

The most straightforward form of single-pixel imaging involves raster scanning, where a single-pixel detector sequentially measures light intensity at each pixel location, requiring N^2 measurements for an $N \times N$ image. In contrast, *ghost imaging* enables image reconstruction from significantly fewer measurements by illuminating the object (x) with structured light patterns (A) and using a single-pixel detector to measure the total transmitted or reflected intensity [20, 21].

The *sampling ratio* is defined as:

$$\beta = \frac{M}{N^2}, \tag{1}$$

where M is the number of structured light patterns used. Each projected pattern ($A^{(m)}$) is modulated by the object and measured as a scalar intensity, commonly referred to as the bucket sum:

$$b^{(m)} = \sum_{i=1}^{N} \sum_{j=1}^{N} A_{i,j}^{(m)} \cdot x_{i,j}. \tag{2}$$

Early methods, such as Differential Ghost Imaging (DGI), used the bucket detector signals to compute weighted sums of the illumination patterns for object reconstruction, but these approaches often produced low-fidelity results [22]. More advanced methods recast ghost imaging as a linear system:

$$\mathbf{b} = \Psi \mathbf{x}, \tag{3}$$

where Ψ is the sensing matrix formed by flattening and stacking the illumination patterns, and $\mathbf{x}$ is the vectorized image. In this form, a standard solution using the Moore–Penrose pseudoinverse (PI) is given by:

$$\mathbf{x} = \Psi^{\dagger} \mathbf{b}. \tag{4}$$

To enhance reconstruction quality, compressed sensing techniques leverage transform-domain sparsity [23, 24, 25], often employing ℓ_1 or ℓ_2 regularization, resulting in

$$\mathbf{b} = \Psi\Phi\boldsymbol{\alpha}, \quad \text{with } \mathbf{x} = \Phi\boldsymbol{\alpha}. \tag{5}$$

Here, Φ is a sparsifying basis (e.g., DCT or wavelets), and $\boldsymbol{\alpha}$ is a sparse coefficient vector. Solutions are obtained using iterative optimization methods such as Iterative Hard Thresholding (IHT) [23], Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [24], or Alternating Direction Method of Multipliers (ADMM) [25]. Definitions of the variables used above are provided in Appendix A.

Deep neural networks and generative models have been used to improve image reconstruction fidelity [26]. One common approach first applies a classical reconstruction algorithm to produce a low-quality image, which is then enhanced using a trained neural network—such as a CNN, U-Net, or, more recently, a diffusion model [27] [28, 29]. This method benefits from the ability to incorporate learned priors for denoising and super-resolution. However, its performance depends on the quality as well as speed of the initial reconstruction and requires large datasets of paired examples. Another approach uses an end-to-end strategy by embedding elements of the ghost imaging process directly into the neural network architecture [30, 31]. This method is typically faster, as it bypasses classical reconstruction algorithms entirely. However, prior studies have been constrained to low-resolution images (e.g., 28×28) and relied on binary mask patterns.

A critical oversight in most work on ghost imaging is the assumption of simultaneous pattern acquisition. This ignores the slow, serial nature of real-world SLM/DMD-based systems, which makes dynamic imaging impractical. In contrast, our hardware-software co-design uses dual-comb interferometry to achieve true parallel acquisition. This makes high-fidelity imaging of fast-moving objects feasible and aligns the physical system with the assumptions of modern reconstruction algorithms.

3 Dual-Comb Ghost Imaging: Experimental Details

Figure 1 illustrates the concept of dual-comb ghost imaging. An optical frequency comb is a laser source whose spectrum consists of a series of equally space, mutually coherent narrow frequency lines, resembling the teeth of a comb. In our approach, a set of 2D speckle patterns H, comprising uncorrelated speckle distributions across different comb line frequencies, is generated at the fiber tip using dual optical frequency combs (OFCs) and projected onto the 2D target object x. The encoded intensity distribution ($H \times x$) is detected by a single-pixel photodetector, producing a dual-comb interferogram. A fast Fourier transform (FFT) converts this time-domain signal into frequency-domain bucket-sum data $y = H \times x$. By multiplexing light through the optical fiber, we parallelize the speckle projection process in ghost imaging, significantly increasing imaging speed.

For the dual-comb OFC source, we use two electro-optic (EO) combs with slightly different free spectral ranges (f_{FSR}). A continuous-wave (CW) fiber laser at $1550\,\text{nm}$ is first amplified and split into two beams via a 50/50 fiber coupler. Each beam is frequency-shifted using an acousto-optic modulator, introducing a center frequency offset Δf_{center}. These beams are then independently modulated by resonant EO modulators driven at f_{FSR} and $f_{\text{FSR}} + \Delta f_{\text{FSR}}$, respectively, where $f_{\text{FSR}} = 20\,\text{GHz}$ and $\Delta f_{\text{FSR}} = 200\,\text{Hz}$. The resulting EO combs are recombined using a 50/50 fiber coupler and amplified to compensate for insertion losses. When two mutually coherent combs with slightly different repetition rates are combined and photodetected, their interference produces a set of beat signals in the radio-frequency (RF) domain. This allows fast, precise electronic detection without slow optical spectrum analyzers, while mutual coherence enables coherent averaging for improved signal-to-noise ratio (SNR) [17].

Before the free-space imaging setup (see Figure 1b), a Waveshaper is placed in the system to filter individual comb lines if needed. For bucket-sum measurements, the Waveshaper transmits the full dual-comb spectrum. The multimode fiber used in the experiments has a core diameter of 200 um and exhibits hundreds of core modes, resulting in speckle patterns when light is coupled to the core of the optical fiber. The generated speckle patterns are then collimated and passed through the target object. A 50/50 beam splitter divides the beam into signal and reference paths, allowing simultaneous acquisition and suppression of common-mode temporal noise. Calibration measurements are performed without the target to correct for optical path imbalances. The set of speckle patterns (H) without the target object is separately recorded using a 2D camera by selecting individual comb lines with the Waveshaper, either before or after the imaging. The imaging target is a

negative USAF 1951 resolution chart mounted on a motorized stage. Bucket-sum signals are acquired using 500 kHz bandwidth free-space InGaAs photodetectors, and speckle patterns are recorded with a 256×256 pixel InGaAs camera.

The imaging experiment uses approximately 200 comb lines spanning from 1530 nm to 1565 nm. The speckle patterns at different frequencies are uncorrelated, as indicated by low Pearson correlation coefficients (Figure 2b), with slight residual correlations attributed to background contributions at the pattern edges. Power variations among comb lines are normalized during processing. Figure 2c shows the signal and reference RF combs obtained via FFT of the interferograms. The amplitude ratios of corresponding comb peaks represent the bucket-sum signals that encode the image information. A calibration measurement without the target is used to correct for the path differences.

For image reconstruction, we used the calibrated bucket-sum signals and the measured speckle patterns. Figure 2d shows the calibrated measurement compared to the theoretical predictions, showing excellent agreement. Images are initially reconstructed using the Moore-Penrose pseudoinverse, recovering the target image even at a 0.3% sampling ratio. However, background noise from speckle structures remains significant. To address this, we developed Optical Ghost-GPT, a transformer-based reconstruction algorithm that substantially suppresses noise and enhances image quality. Details of Optical Ghost-GPT will be discussed in the following section.

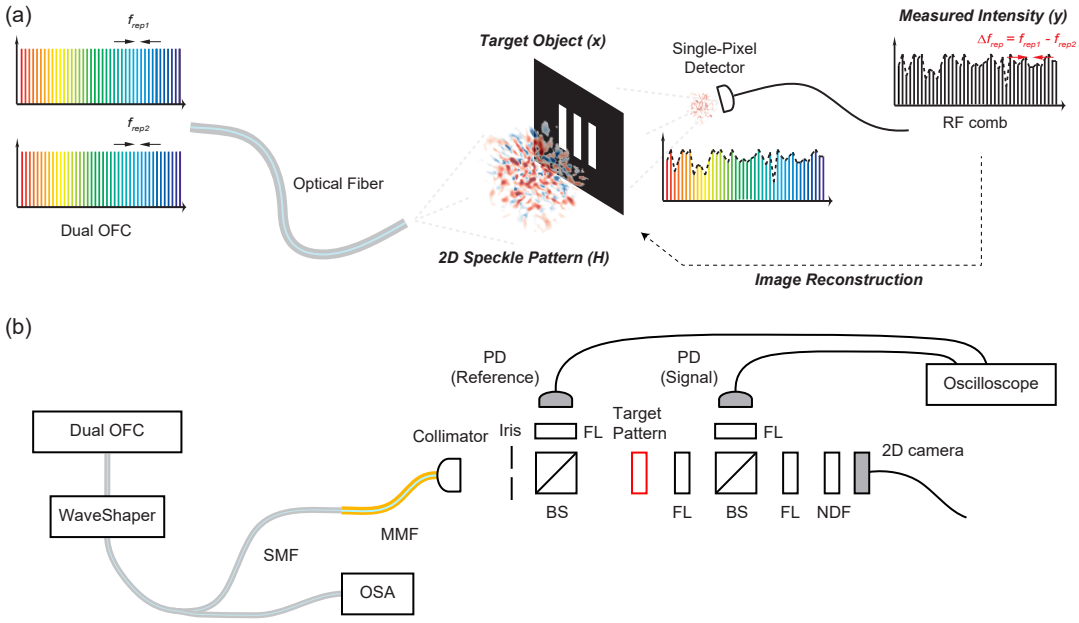


Figure 1: **Optical Fiber-based Dual-Comb Ghost Imaging.** (a) Conceptual illustration of optical fiber-based ghost imaging using dual optical frequency combs. Each comb line generates an uncorrelated speckle pattern at the fiber tip, and the set of speckle patterns is mapped onto the target object x. The encoded speckle pattern ($H \times x$) is collected by a single-pixel detector, and the resulting dual-comb interference signal ("interferogram") is recorded. The time-domain interferogram is converted into the frequency domain by fast Fourier transform (FFT), yielding the bucket-sum information ($y = H \times x$). The dual-comb technique allows for snapshot ghost imaging, providing a highly parallel, high-SNR, and fast imaging capability. Notably, the use of comb-based wavelength-division multiplexing (WDM) and bucket-sum compressive imaging reduces the physical dimensions of both input and output hardware to essentially zero-dimensional (i.e., single spot or pixel), making the system particularly suitable for applications requiring compact form factors, such as endomicroscopy. (b) Experimental setup diagram. SMF: Single Mode Fiber, OSA: Optical Spectrum Analyzer, PD: Photodetector, FL: Focusing Lens, BS: Beam Splitter, NDF: Neutral Density Filter.

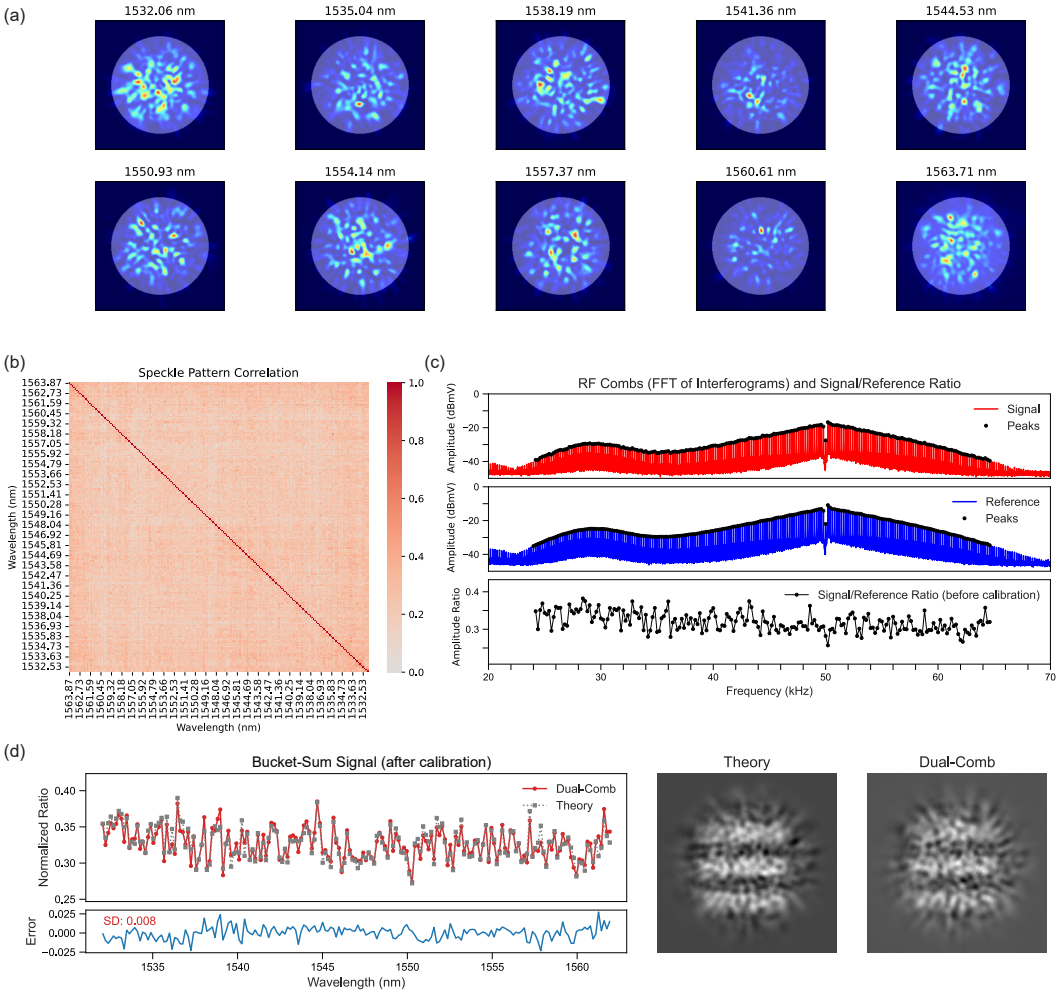


Figure 2: **Experimental Details.** (a) Speckle patterns measured at ten different comb line frequencies. Each image contains 256×256 pixels. (b) Pearson correlation coefficients calculated between speckle patterns across the wavelength range from 1532 nm to 1564 nm. The correlation is computed within the bright circular region of each speckle pattern. The low off-diagonal values indicate that the speckle patterns are largely uncorrelated. Residual correlations are attributed primarily to the common background near the edges of the speckle patterns. (c) Radio-frequency (RF) combs generated from dual-comb interference signals measured by the signal and reference single-pixel detectors (upper and middle panels). The ratio between the comb peak amplitudes of the signal and reference RF combs is shown (bottom panel). Differences between the signal and reference beam paths are calibrated using measurements performed with and without the target object. (d) The calibrated bucket-sum signal measured from the dual-comb experiment shows excellent agreement with the theoretical prediction, which is obtained by masking the ground-truth pattern onto the stored speckle pattern images. The standard deviation (SD) between the two bucket-sum signals is 0.008. Image reconstruction is performed using the Moore-Penrose pseudoinverse algorithm at a sampling ratio (SR) of 0.29%. Both the theoretical reconstruction (left) and the experimental dual-comb reconstruction (right) successfully recover the ground-truth patterns.

4 Transformer Modeling

Transformers are a powerful deep learning architecture originally introduced for natural language processing (NLP) tasks but have since found applications in various domains [32], including computer vision [33], speech processing [34], and scientific data analysis [35]. They leverage a mechanism called self-attention to model long-range dependencies and capture contextual relationships within sequences [36]. Unlike traditional recurrent or convolutional networks, transformers process entire sequences simultaneously, making them highly efficient for parallel computation. Their ability to learn complex patterns from large datasets has made them the backbone of state-of-the-art models like BERT, GPT [37], and Vision Transformers (ViTs) [38] [39].

In typical ViTs, the image is broken up into equally sized patches, which serve as the ViTs' token. In order to adapt this methodology to dual-comb ghost imaging, we propose concatenating the flattened illumination patterns that make the sensing matrix, Ψ, with the bucket value to form the token for our model.

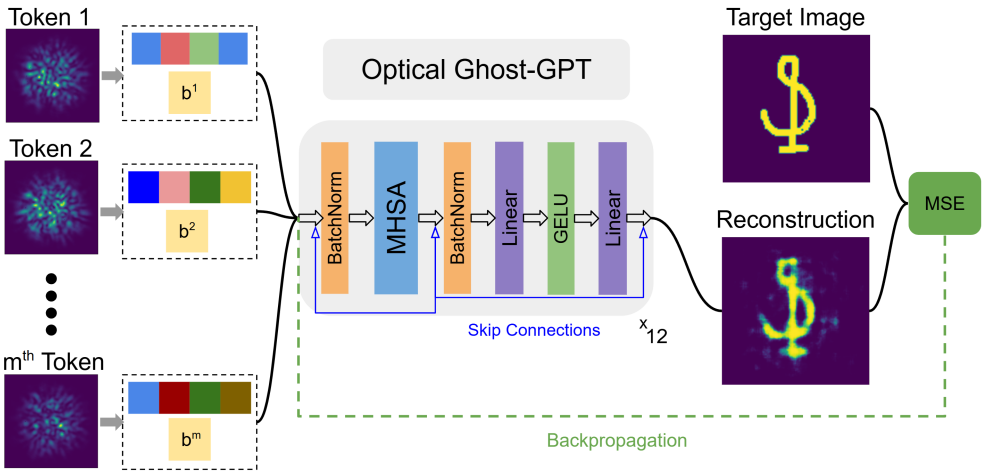


Figure 3: **Schematic of Optical Ghost-GPT**. The speckle patterns are first compressed into a latent space and then concatenated with the corresponding bucket measurements to form the input token sequence.

Model Architecture. We introduce **Optical Ghost-GPT** (or simply **Ghost-GPT**), a transformer-based model designed for structured image reconstruction from ghost imaging measurements. The model leverages stacked self-attention mechanisms and residual feedforward blocks to model long-range dependencies across the contextual input.

Input Embedding. The model receives two primary inputs in each token: a flattened image of the speckle pattern $\mathbf{\Psi}^m \in \mathbb{R}^N$, where $N = 256 \times 256$, and bucket sum value $\mathbf{b^m}$. The image vector is projected via a learnable linear transformation to a latent embedding of dimension `embedding_dim` $-$ 1, and afterwards, the latent representation of the image vector and its corresponding bucket sum are concatenated. To encode positional structure, we add learned positional embeddings of size `embedding_dim` to each token in the sequence:

$$\mathbf{z}_i = \mathbf{e}_i + \mathbf{p}_i, \quad i = 1, \ldots, C,$$

where $\mathbf{e}_i$ is the input token embedding, $\mathbf{p}_i$ is the corresponding positional embedding, and C=250 is the context size determined by the theorectical maximum number of RF comblines in our setup.

Transformer Blocks. The architecture contains a stack of $L = 12$ transformer blocks, each composed of a multi-head self-attention (MHSA) mechanism and a two-layer feedforward network. The attention mechanism employs H number of heads, computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where Q, K, and V are linear projections of the input sequence, and $d_k = \texttt{embedding\_dim}$. We also apply a dropout mask with a value of 0.1 to ensure robustness of the model and allow the model to generalize to missing shots and buckets. Each block includes residual connections and a batch normalization operation placed before and after the MHSA layer, defined as:

$$\text{BatchNorm}(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \quad \mu = \mathbb{E}[\mathbf{x}], \quad \sigma^2 = \text{Var}[\mathbf{x}].$$

The feedforward network consists of two linear transformations with a gaussian error linear unit (GELU) activation; A final transformer block is appended after the main stack to further refine the sequence representation.

$$\text{FFN}(\mathbf{x}) = \text{Linear}_2\left(\text{GELU}(\text{Linear}_1(\mathbf{x}))\right).$$

Output Projection. The final token representations are projected to $\mathbb{R}^{16}$ via a linear layer and the output is flattened to a tensor of size $(C \times 16)$ and passed through a final linear layer to reconstruct the original image vector in $\mathbb{R}^{256\times256}$. A sigmoid activation is applied to constrain the output to the range $[0, 1]$, consistent with normalized image intensities:

$$\hat{\mathbf{x}} = \sigma\left(\text{Linear}_{\text{final}}(\text{Flatten}(\mathbf{z}))\right).$$

Network Training. We first obtain a set of speckle patterns from the 2-D camera during the calibration phase of the experiment. (We emphasize that the 2-D camera is only needed to obtain the initial speckle patterns and can be removed during imaging). For this experiment, we obtained 188 speckle patterns, which are then used to generate synthetic buckets sums via a convolution between the digitized speckle pattern and images from the MNIST and OMNIglot datasets. We form our labeled dataset of synthetic bucket sums as the x-label and its corresponding target images as the y-label. (See Appendix B for train/test split and computational resources used).

In training, Ghost-GPT predicts the target image, given the speckle pattern and bucket sums as the input. We use mean squared error as our loss function and an AdamW optimizer with a learning rate of 0.0003 and a weight decay of 0.001. For the following reconstruction results, we used a model with the number of attention heads set to 8 and the `embedding_dim` set to 32 based on a hyperparameter sweep. (See Appendix C for further information on the hyperparameter analysis).

5 Ghost-GPT Reconstruction Results

In this section, we examine the reconstruction results in simulation using our experimental speckle pattern results. Figure 4 shows a series of reconstruction compared with their true images.

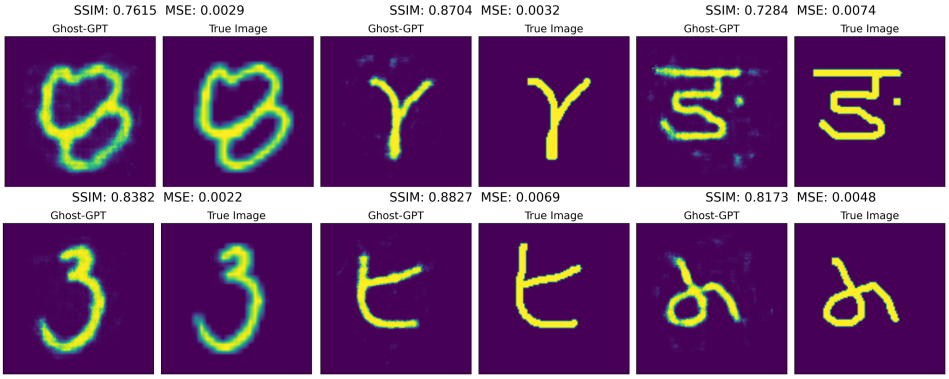


Figure 4: Reconstructions from Ghost-GPT versus the true image with the Structural Similarity Index Measure (SSIM) and Mean Squared Error (MSE) displayed above.

We emphasize that we achieve very high SSIM (greater than 0.7) and low MSE (less than 0.02) values despite having a sampling ratio of only 0.29%. We observe that the model is able to preserve fine structural details and clear object boundaries, highlighting the model's ability to recover meaningful

image content from minimal measurements. However, the reconstructions can exhibit variations in intensity across the object, especially in images with long, thin structures or uniform intensity profiles. These artifacts likely stem from the inherent non-uniformity of the speckle patterns generated by the multi-mode fiber. This issue could potentially be mitigated by employing a more uniform speckle distribution or by including a smoothness term, such as total variational loss, in the loss function.

We also compared our model against classical reconstruction algorithms on 256 images from our validation dataset. (256 images were chosen due to the long reconstruction times associated with the iterative FISTA algorithm). Table 1 compares the MSE and SSIM of previously discussed reconstruction algorithms- as expected the simpler reconstruction algorithms such as Differential Ghost Imaging and the Moore-Penrose Psuedo Inverse perform worse than compressed sensing methods like FISTA. However, Ghost-GPT outperforms these classical algorithms giving an average **MSE of 0.008** and **SSIM of 0.788**, while being approximately **263x** faster than FISTA. (We set $\epsilon = 50$ after performing a hyperparameter sweep and choosing the best performing SSIM). The extremely fast reconstruction speed of **14 ms** enables **real-time, video frame-rate ghost imaging in optical fibers**. Importantly, while the computational reconstruction speed can be further improved with better computing hardware, the fundamental limit of image reconstruction is set by the repetition rate difference of the dual-comb, which is typically a few hundred Hz to several kHz in our experiments. With a larger repetition rate difference, a much higher frame rate is possible. However, this requires high-bandwidth photodetection, and the trade-off between the sampling ratio and frame rate must be considered in accordance with the Nyquist condition.

Table 1: Classical Algorithms vs Ghost-GPT

Algorithm	MSE	SSIM	Computational Time (ms)
Differential Ghost Imaging	0.184 (σ=0.037)	0.042 (σ=0.022)	15.7 (σ=0.138) GPU
Moore–Penrose Pseudo-Inverse	0.093 (σ=0.027)	0.055 (σ=0.027)	5640 (σ=52.1) CPU 157 (σ=3.27) GPU
FISTA (200 iters, ϵ=50)	0.045 (σ=0.017)	0.092 (σ=0.019)	3680 (σ=125) CPU 10500 (σ=93.8) GPU
Ghost-GPT (Ours)	**0.008 (σ=0.009)**	**0.788 (σ=0.076)**	**14.0 (σ=0.457) GPU**

With our experimental setup, we collected physical intensity measurements of a USAF resolution test target (ThorLabs R3L3S1N 3" x 3"). Figure 5 and Table 2 shows the reconstruction results of a stripe (top) and the number 2 (bottom) compared with classical algorithms and Ghost-GPT and the corresponding MSE and SSIM values. In real world environments, we verify that Ghost-GPT is able to successfully reconstruct images with higher fidelity for both targets compared to classical methods.

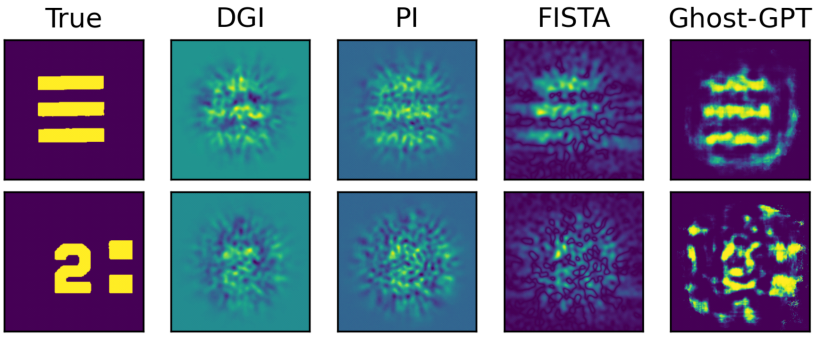


Figure 5: Reconstruction using experimental bucket measurements compared with classical algorithms and Ghost-GPT. (Top row): Striped lines correspond to group 0, element 4 (1.41 line pairs per millimeter). (Bottom Row): The number 2 corresponds to group 0, element 2 (1.12 lp/mm).

6 Robustness of Ghost-GPT in Experimental Imaging

To examine the effect of noise in our experimental measurement, we performed a signal-to-noise ratio analysis by adding varying amounts of noise to simulated buckets and gauged the quality

Table 2: MSE/SSIM for Experimental Targets with Classical Algorithms and Ghost-GPT

Experimental Target	DGI	PI	FISTA	Ghost-GPT (Ours)
Stripes	0.231/0.030	0.140/0.028	0.072/0.057	**0.058/0.470**
Number 2	0.204/0.036	0.138/0.028	0.084/0.064	**0.077/0.484**

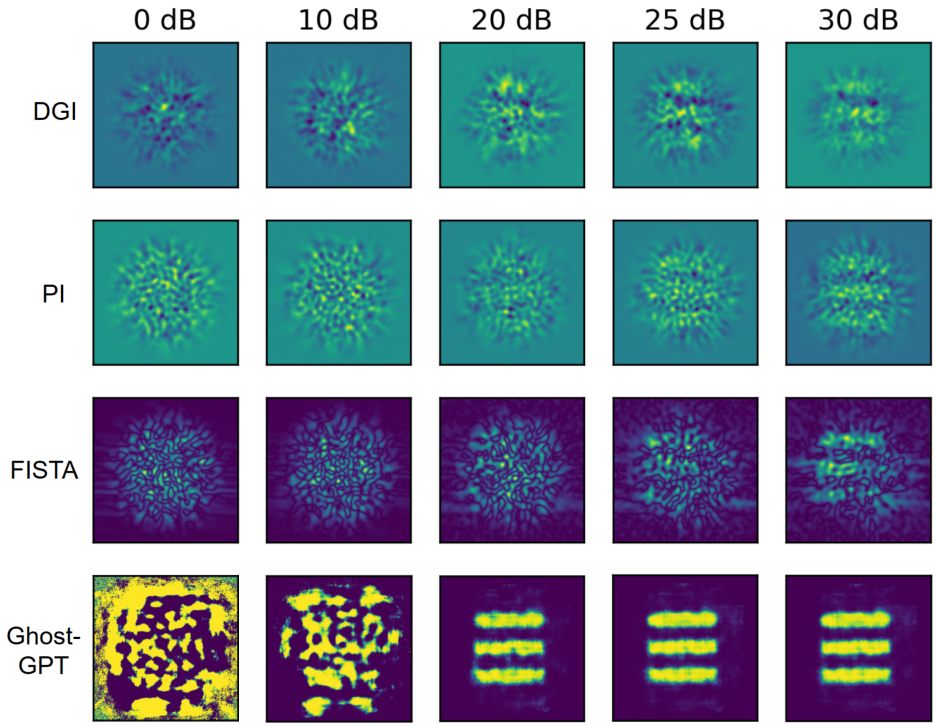


Figure 6: Simulated reconstructions of the USAF striped targets with artificial noise added to the buckets. The signal to noise ratio in dB is shown label at the top with the corresponding reconstruction algorithms on the left-hand side.

of image reconstructions (Figure 7). The artificial noise added to the buckets were calculated by sampling points from a gaussian with mean of zero and standard deviation of one and scaling it inversely proportional to the value of the SNR. (See Appendix D.1 for more information about the SNR calculation). Ghost-GPT demonstrates strong robustness to noise, outperforming classical reconstruction algorithms, with recognizable reconstructions achievable at SNR levels above 20 dB for Ghost-GPT. (See Appendix Figure 9 for plots of the calculated MSE and SSIM values vs SNR). Based on this analysis, our experimental bucket measurements are approximately equivalent to 20-25 dB. For further experimental results, code/data availability, and discussion about project limitations, please refer to D.2, E, and G.

7 Comparison with Other Deep Learning Models

We compared Ghost-GPT, our transformer-based reconstruction model, against two deep learning baselines: a decoder-only CNN and a U-Net with skip connections. All models were matched in parameter count ($\sim$ 270M) and trained on identical speckle–bucket inputs for fair comparison. As shown in Table 3, Ghost-GPT achieves the lowest mean-squared error (MSE = 0.0068), representing a 47.7% reduction vs. CNN and 38.18% vs. U-Net across the entire validation dataset, while maintaining comparable structural similarity (SSIM = 0.787).

Training dynamics further show that Ghost-GPT converges to a lower validation loss within the same number of epochs, supporting the suitability of the transformer's self-attention mechanism

Table 3: Image Reconstruction Quality on Various Neural Network Architectures

Model Architecture	MSE	SSIM	Computational Time (ms)
Ghost-GPT (Ours)	**0.0068 (σ=0.0077)**	0.787 (σ=0.073)	14.0 (σ=0.457) (GPU)
CNN	0.013 (σ=0.0086)	0.8079 (σ=0.0395)	**6.26 (σ=0.41) (GPU)**
UNet	0.011 (σ=0.0073)	**0.8316 (σ=0.0696)**	15.09 (σ=0.032) (GPU)

for modeling the global correlations inherent in ghost imaging measurements.These findings align with recent studies reporting transformer models outperforming conventional deep architectures in computational imaging tasks [40]. While our model was trained using MSE loss for quantitative comparability, the observed SSIM gap suggests that training with a perceptual loss or SSIM-weighted objective could further enhance structural fidelity without sacrificing reconstruction accuracy. We leave this as a promising direction for improving perceptual quality in future iterations of Ghost-GPT. It is worth noting that, with simple binary datasets, it was difficult to compare GhostGPT to other deep-learning models because the performance metrics gave mixed signals within a narrow range. With a larger number of speckle patterns and more realistic grayscale datasets, we should be able to conduct a more systematic performance comparison with other deep-learning models.

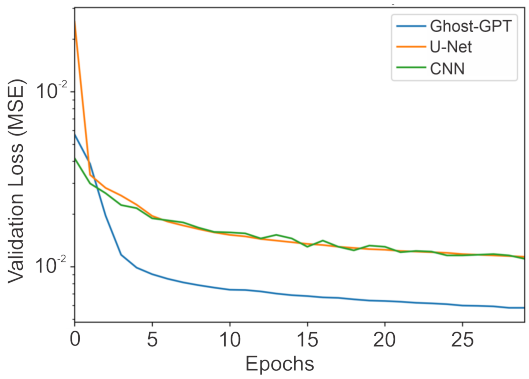


Figure 7: Loss history of different deep learning models compared with Ghost-GPT. Our transformer-based model consistently outperforms both U-Net and CNN for large majority of epochs.

8 Conclusion

In this work, we demonstrate a novel dual-comb ghost imaging using only a single-core fiber and a single-pixel detector by combining dual-comb interferometry with a Transformer tailored for fast, high-quality reconstruction—a hardware–software co-design. Our GPT-like deep learning model is application-aware: paired tokens jointly encode each speckle pattern and its corresponding bucket sum to meet real-time, high-fidelity goals. Leveraging such deep learning architectures, we achieved high-fidelity reconstructions (average SSIM of 0.788) at ultra-low sampling ratios (as low as 0.29%), significantly outperforming classical methods in both resolution and efficiency. In our first experimental demonstration, the number of uncorrelated speckle patterns was limited to about 200, so we used binary datasets suitable for reconstruction at such low sampling ratios. This, however, is a prototype-stage constraint, not a fundamental limitation of the method. Realistic grayscale reconstructions could be enabled by increasing the number of comb lines, which can be achieved by broadening the spectrum and reducing speckle correlation width. Future work will focus on achieving video-rate grayscale imaging, enhancing the model via improved sensing matrix design and noise-aware training, and developing efficient beam collection in reflective-mode setups for practical applications. This framework holds strong potential for emerging applications in biomedical imaging and imaging within extreme, low-light environments. The ability to reconstruct high-resolution images from sparse, low-light measurements makes it particularly suited for imaging live or light-sensitive biological samples, for real-time, minimally invasive procedures such as fiber-based endoscopy, or dynamic tissue monitoring—offering a promising pathway toward next-generation medical diagnostics.

Acknowledgments and Disclosure of Funding

DD, MGS, and HWHL conceived the concept of Optical Ghost-GPT. DD and WJMKK designed and implemented the GPT model architecture. DD, BH, and MGS carried out model training, evaluation, and comparative analysis with classical reconstruction algorithms. MGS, MG, and BJP led the experimental integration of dual-comb spectroscopy with optical fiber-based ghost imaging. YJ explored various optical fiber types for this application and conducted speckle pattern simulations using Finite Difference Time Domain (FDTD) methods. All authors contributed to the discussions and writing of the manuscript.

DD acknowledges the support by UCI-LANL-SoCal Hub graduate fellowship program. WJMKK acknowledges the LANL Laboratory Directed Research and Development program for funding under project 20250492ER.

References

[1] A. Boese, C. Wex, R. Croner, U. B. Liehr, J. J. Wendler, J. Weigt, T. Walles, U. Vorwerk, C. H. Lohmann, M. Friebe, *et al.*, "Endoscopic imaging technology today," *Diagnostics*, vol. 12, no. 5, p. 1262, 2022.

[2] S. Urayama, R. Kozarek, and S. Raltz, "Evaluation of per-procedure equipment costs in an outpatient endoscopy center," *Gastrointestinal endoscopy*, vol. 44, no. 2, pp. 129–132, 1996.

[3] A. F. Gmitro and D. Aziz, "Confocal microscopy through a fiber-optic imaging bundle," *Optics letters*, vol. 18, no. 8, pp. 565–567, 1993.

[4] B. A. Flusberg, E. D. Cocker, W. Piyawattanametha, J. C. Jung, E. L. Cheung, and M. J. Schnitzer, "Fiber-optic fluorescence imaging," *Nature methods*, vol. 2, no. 12, pp. 941–950, 2005.

[5] C. M. Lee, C. J. Engelbrecht, T. D. Soper, F. Helmchen, and E. J. Seibel, "Scanning fiber endoscopy with highly flexible, 1 mm catheterscopes for wide-field, full-color imaging," *Journal of biophotonics*, vol. 3, no. 5-6, pp. 385–407, 2010.

[6] W. Choi, M. Kang, J. H. Hong, O. Katz, B. Lee, G. H. Kim, Y. Choi, and W. Choi, "Flexible-type ultrathin holographic endoscope for microscopic imaging of unstained biological tissues," *Nature communications*, vol. 13, no. 1, p. 4469, 2022.

[7] J. Sun, R. Kuschmierz, O. Katz, N. Koukourakis, and J. W. Czarske, "Lensless fiber endomicroscopy in biomedicine," *PhotoniX*, vol. 5, no. 1, p. 18, 2024.

[8] C. Liu, J. Chen, J. Liu, and X. Han, "High frame-rate computational ghost imaging system using an optical fiber phased array and a low-pixel apd array," *Optics express*, vol. 26, no. 8, pp. 10048–10064, 2018.

[9] M. J. Padgett and R. W. Boyd, "An introduction to ghost imaging: quantum and classical," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 375, no. 2099, p. 20160233, 2017.

[10] D. Li, D. Yang, S. Sun, Y.-G. Li, L. Jiang, H.-Z. Lin, and W.-T. Liu, "Enhancing robustness of ghost imaging against environment noise via cross-correlation in time domain," *Optics Express*, vol. 29, no. 20, pp. 31068–31077, 2021.

[11] X. Liu, J. Shi, L. Sun, Y. Li, J. Fan, and G. Zeng, "Photon-limited single-pixel imaging," *Optics express*, vol. 28, no. 6, pp. 8132–8144, 2020.

[12] X. Liu, J. Shi, X. Wu, and G. Zeng, "Fast first-photon ghost imaging," *Scientific reports*, vol. 8, no. 1, p. 5012, 2018.

[13] F. Li, M. Zhao, Z. Tian, F. Willomitzer, and O. Cossairt, "Compressive ghost imaging through scattering media with deep learning," *Optics Express*, vol. 28, no. 12, pp. 17395–17408, 2020.

[14] L.-X. Lin, J. Cao, D. Zhou, H. Cui, and Q. Hao, "Ghost imaging through scattering medium by utilizing scattered light," *Optics Express*, vol. 30, no. 7, pp. 11243–11253, 2022.

[15] M. Don, "An introduction to computational ghost imaging with example code," *CCDC Army Research Laboratory: Aberdeen Proving Ground, MD, USA*, 2019.

[16] V. Kilic, T. D. Tran, and M. A. Foster, "Compressed sensing in photonics: tutorial," *Journal of the Optical Society of America B*, vol. 40, no. 1, pp. 28–52, 2022.

[17] I. Coddington, N. Newbury, and W. Swann, "Dual-comb spectroscopy," *Optica*, vol. 3, no. 4, pp. 414–426, 2016.

[18] C. Bao, M.-G. Suh, and K. Vahala, "Microresonator soliton dual-comb imaging," *Optica*, vol. 6, no. 9, pp. 1110–1116, 2019.

[19] E. Vicentini, Z. Wang, K. Van Gasse, T. W. Hänsch, and N. Picqué, "Dual-comb hyperspectral digital holography," *Nature Photonics*, vol. 15, no. 12, pp. 890–894, 2021.

[20] S. Li, F. Cropp, K. Kabra, T. Lane, G. Wetzstein, P. Musumeci, and D. Ratner, "Electron ghost imaging," *Physical review letters*, vol. 121, no. 11, p. 114801, 2018.

[21] Y. Bromberg, O. Katz, and Y. Silberberg, "Ghost imaging with a single detector," *Phys. Rev. A*, vol. 79, p. 053840, May 2009.

[22] H.-C. Liu, "Imaging reconstruction comparison of different ghost imaging algorithms," *Scientific Reports*, vol. 10, no. 1, p. 14626, 2020.

[23] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[24] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[25] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[26] S. Rizvi, J. Cao, K. Zhang, and Q. Hao, "Deepghost: real-time computational ghost imaging via deep learning," *Scientific reports*, vol. 10, no. 1, p. 11400, 2020.

[27] Y. He, G. Wang, G. Dong, S. Zhu, H. Chen, A. Zhang, and Z. Xu, "Ghost imaging based on deep learning," *Scientific reports*, vol. 8, no. 1, p. 6469, 2018.

[28] F. Wang, C. Wang, M. Chen, W. Gong, Y. Zhang, S. Han, and G. Situ, "Far-field super-resolution ghost imaging with a deep neural network constraint," *Light: Science & Applications*, vol. 11, no. 1, p. 1, 2022.

[29] S. Mao, Y. He, H. Chen, H. Zheng, J. Liu, Y. Yuan, M. Le, B. Li, J. Chen, and Z. Xu, "High-quality and high-diversity conditionally generative ghost imaging based on denoising diffusion probabilistic model," *Optics Express*, vol. 31, no. 15, pp. 25104–25116, 2023.

[30] W. Ren, X. Nie, T. Peng, and M. O. Scully, "Ghost translation: an end-to-end ghost imaging approach based on the transformer network," *Optics Express*, vol. 30, no. 26, pp. 47921–47932, 2022.

[31] J. Liang, Y. Cheng, and J. He, "Transformer-based flexible sampling ratio compressed ghost imaging," *Engineering Analysis with Boundary Elements*, vol. 170, p. 106050, 2025.

[32] D. Rothman, *Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4*. Packt Publishing Ltd, 2022.

[33] S. Jamil, M. Jalil Piran, and O.-J. Kwon, "A comprehensive survey of transformers for computer vision," *Drones*, vol. 7, no. 5, p. 287, 2023.

[34] S. Kashyap, S. Singh, and D. V. Singh, "Speech-to-speech translation using transformer neural network," in *International conference on soft computing for problem-solving*, pp. 813–826, Springer, 2023.

[35] F. Sufi, "Generative pre-trained transformer (gpt) in research: A systematic review on data augmentation," *Information*, vol. 15, no. 2, p. 99, 2024.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[37] B. Ghojogh and A. Ghodsi, "Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey." working paper or preprint, Dec. 2020.

[38] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences*, vol. 13, no. 9, p. 5521, 2023.

[39] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2998–3008, 2021.

[40] Y. Chen, H. An, Z. Sun, T. Tian, M. Chen, C. Spielmann, and X. Li, "Large model enhanced computational ghost imaging," 2025.

[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[42] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," in *Science*, vol. 350, pp. 1332–1338, American Association for the Advancement of Science, 2015.

A Variable Definitions

The equations in the main text use the following symbols:

- $A^{(m)} \in \mathbb{R}^{N \times N}$: The m^{th} structured illumination pattern,
- $x \in \mathbb{R}^{N \times N}$: The unknown image to be reconstructed,
- $b^{(m)}$: Total detected intensity for the m^{th} pattern,
- $\mathbf{b} \in \mathbb{R}^{M \times 1}$: Vector of all scalar intensity measurements,
- $\Psi \in \mathbb{R}^{M \times N^2}$: Sensing matrix composed of flattened patterns,
- $\mathbf{x} \in \mathbb{R}^{N^2 \times 1}$: Flattened version of the image x,
- $\Phi \in \mathbb{R}^{N^2 \times N^2}$: Sparsifying basis (e.g., DCT, wavelets),
- $\boldsymbol{\alpha} \in \mathbb{R}^{N^2 \times 1}$: Sparse coefficients in the transform domain.

B Dataset Generation, Train and Test Split

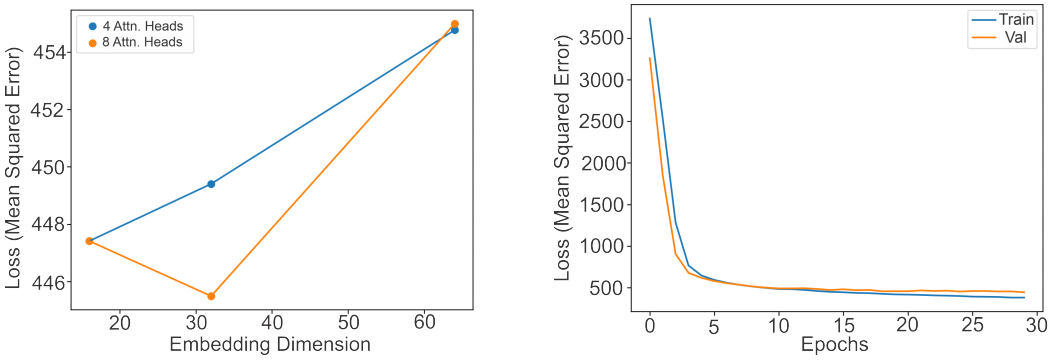
Since the bucket signal is effectively a convolution between the object and each speckle pattern, it is computationally efficient to synthetically generate labeled training data by pairing known objects with their corresponding bucket sums. For this, we use open-source image datasets and simulate the measurement process. Our dataset includes 19,280 from OMNIglot and 19,280 images from MNIST with a train/val split of 33,000/5,560 respectively. All images are resized to 256 × 256 to match the resolution of the recorded speckle patterns. We use a batch size of 32 for the training dataset and a batch size of 64 for the validation dataset.

B.1 Licensing

MNIST: Originally published by [41], downloaded via `torchvision.datasets.MNIST`, and licensed under the *Creative Commons Attribution-Share Alike 3.0* license (see `http://yann.lecun.com/exdb/mnist/`).

Omniglot: Originally published in [42], downloaded via `torchvision.datasets.Omniglot`, and licensed under the *MIT License* (see `https://github.com/brendenlake/omniglot`).

C Hyperparameter Analysis



(a) (b)

Figure 8: The lowest loss value on the validation dataset as a function of the number of attention heads and embedding dimensions (a) comparison with pattern embedding (b) log scale loss values of models with and without pattern embedding

For hyperparameter tuning, we performed a simple parameter sweep by changing the number of attention heads at values of 4 and 8 as well as adjusting the embedding dimension at values of 16, 32,

and 64. Due to memory constraints, we limited our parameter search to these values. As previously stated, we use mean squared error as our loss function and an AdamW optimizer with a learning rate of 0.0003 and a weight decay of 0.001. In our loss function, we add the error in each pixel and then average over the batch size.

Over 30 epochs, the model took approximately 1 hour to train on a Linux workstation, using an AMD Ryzen Threadripper 3990X 64-Core Processor and four NVIDIA A6000 GPUs. Our chosen model (8 attention heads, embedding dimension 32) contained $\sim$ 270 million trainable parameters and demonstrates convergence afters 10 epochs (Appendix Figure 8b)- displaying no characteristics of overfitting.

D Further Experimental Results

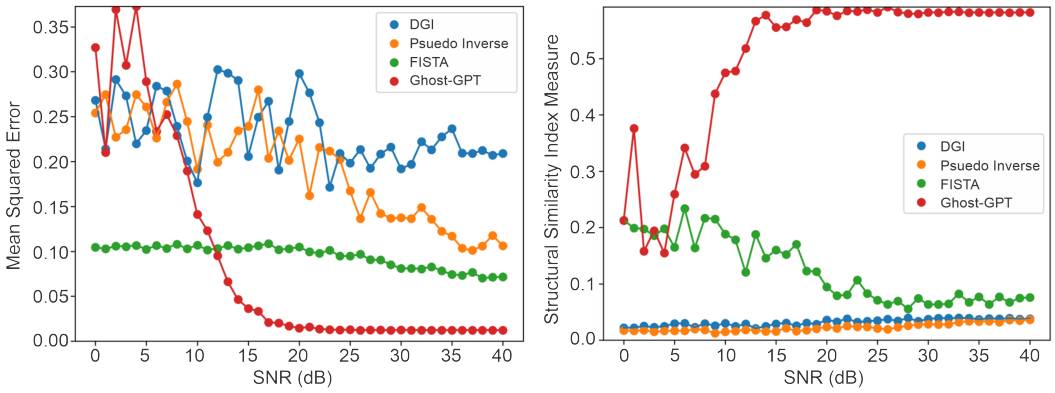
D.1 Different Algorithms vs SNR

The SNR was calculated via the following equation:

$$\mathbf{b}_{noisy} = \mathcal{N}(0,1) * \sigma + \mathbf{b}, \quad \sigma = \frac{\mathbb{E}[\mathbf{b}]}{\text{SNR}^{10}}$$

We sample points from a gaussian of mean 0 and std 1, which is scaled by σ, the standard deviation. σ is calculated the by averaging over the collection of bucket sum and then dividing by the SNR value raised to the 10^{th} power. The true bucket sum, $\mathbf{b}$, is added to the previously calculated quantity to create the noisy bucket, $\mathbf{b}_{\text{noisy}}$. Previous works in ghost imaging formulated their SNR analysis using these exact definitions.

Appendix Figure 9 plots the MSE and SSIM of the reconstructed images compared with classical reconstruction algorithms and Ghost-GPT.



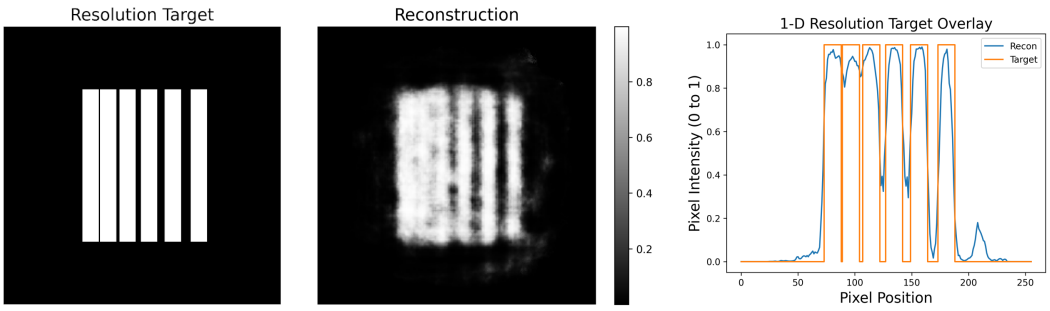
(a) (b)

Figure 9: SNR ratio in the buckets versus (a) MSE and (b) SSIM of the USAF target

D.2 Resolution Analysis

In ghost imaging, resolution is fundamentally linked to the spatial characteristics of the light patterns used to probe the object—specifically, the grain size of these patterns. Grain size refers to the typical scale or correlation length of the intensity fluctuations in the illumination patterns (often speckle patterns). This grain size determines the finest detail that can be distinguished in the reconstructed image.

In this section, we evaluated the smallest resolvable feature size achievable by Ghost-GPT in simulation using our current experimental speckle pattern set (Appendix Figure 10. A digital resolution target was generated with 15-pixel-wide bars and varying gap sizes (1, 3, 5, 7, 9 pixels). Ghost-GPT successfully resolved features with a 5-pixel gap. In our experimental setup, the Air Force resolution target's 23-pixel feature between a pair of stripes corresponds to a physical size of 0.355 mm, indicating that Ghost-GPT achieves a resolution of approximately 0.077 mm. At



(a) (b) (c)

Figure 10: Simulated resolution testing of Ghost-GPT. (a) True resolution target with increasing pixel gaps (b) Ghost-GPT reconstruction results (c) Line profile comparing the ground truth and model output, obtained by scanning along a horizontal cross-section midway through the images.

this resolution, our ghost imaging model is well-suited for visualizing a broad range of biological structures, including tissue architecture such as blood vessels, skin layers, and muscle fibers. It also enables observation of developmental forms in small organisms like embryos and larvae, as well as structural features in plant tissues such as roots and leaves.

E Code Availability

The scripts, specific training/validation dataset, experimental dual-comb speckle patterns, and target measurements are available at: Code Repository Link. Instructions for running the scripts can be found in the `README.txt` file, and the required libraries and dependencies are specified in the `environment.yml` file.

F Broader Impacts

Although there are no evident potential negative societal impacts associated with dual-comb, transformer based ghost imaging, there are many notable positive impacts. These include advances in the health industry, specifically biomedical image quality, speed, and adaptability as clearer images can enable earlier disease detection and more precise surgical guidance. On top of better medical image quality, the integration of deep learning into reconstruction imaging heavily benefits endoscopy devices as it could allow for less invasive fiber-based procedures. In particular, transformer based ghost imaging can also boost overall accuracy by learning and compensating for complex patterns with environmental noise, producing high quality images even in low light or highly scattering conditions.

G Limitations

While Ghost-GPT demonstrates high reconstruction fidelity, computational efficiency, and robustness to noise, there are several limitations with our current work. First, the model performance is contingent on the quality and representativeness of the calibration speckle patterns; any deviations in experimental conditions may lead to distribution shift, potentially reducing reconstruction accuracy. Second, although the model generalizes well to synthetic and experimental data within the resolution constraints of our setup, it has not been benchmarked on more complex, naturalistic scenes or dynamic objects beyond the training domain. Additionally, while the use of synthetic training data from MNIST and Omniglot enables rapid prototyping, the domain gap between these datasets and real biological or clinical targets may necessitate fine-tuning with domain-specific data. Third, our current framework is only comparing our model against classical algorithms.

In this work, we focused our comparison on classical ghost imaging reconstruction algorithms such as Differential Ghost Imaging, the Moore-Penrose pseudoinverse, and FISTA. However, further benchmarking against advanced generative or super-resolution models—such as diffusion models,

adversarial networks, or U-Nets —could offer insight into alternative priors and reconstruction strategies (i.e. more complex loss functions) that may enhance fidelity or reduce artifacts. We emphasize that this work represents an early proof-of-principle for dual-comb ghost imaging using a transformer-based architecture and should be interpreted as a foundation for future improvements. Continued development will include more sophisticated training data, model architectures, and experimental protocols to fully exploit the potential of this novel imaging paradigm.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the paper, we introduced a novel ghost imaging optical setup as well as a transformer based reconstruction model. Our simulated and physical experimental image reconstructions demonstrate high fidelity and fast speeds compared with classical algorithms.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a section in the Appendix, which discusses the limitations of our work (Appendix G).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not contain any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides all the experimental optical equipment used in the results (See Section 3). We describe the model architecture, training, and evaluation in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes] ,

Justification: In Appendix Section E, we provide a link to our code, dataset, and experimental data needed to reproduce the main results in Section 5 and Section 6.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Appendix Section B and C describes the training/validation data splits and hyperparameter used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For the reported MSE and SSIM values of the image reconstructions, we provide the standard deviation in these figures of merits.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Section 3 for the experimental optics setup and Appendix Section C for information regarding the computational resources involved in model training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We completely follow NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential societal impact in Appendix Section F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not use any pretrained language models, image generators, or scraped datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide licensing for the MNIST and OMNIglot datasets used in training for the model.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We attach the link for our code this submission in the appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not conduct any experiments involving human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We used a USAF 1951 target pattern for imaging, and the paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We do not use large language models as a core component of our methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.