Text-Guided Diffusion Model for Adaptive Reconstruction through Scattering Media

Fakhriyya Mammadova [1] **Chanseok Lee** [2] **Mooseok Jang** [2 3]

Abstract

Optical imaging through scattering media presents a fundamentally ill-posed inverse problem, as inhomogeneous materials scramble incident light into seemingly random speckle patterns, destroying phase information and diffraction-limited resolution. While the optical memory effect allows for the recovery of object autocorrelation from scattered intensity, traditional phase-retrieval algorithms routinely stagnate on complex or dense objects. In this paper, we introduce a text-guided diffusion model framework to solve the inverse problem of imaging through scattering layers, leveraging the strong generative priors of diffusion models and the semantic disambiguation of text conditioning. We evaluate our method on the MNIST and Quick, Draw! datasets using experimentally obtained speckle measurements through an optical diffuser. Our results demonstrate that textual prompts effectively resolve the inherent ambiguities of phase loss, significantly outperforming traditional correlation-based imaging techniques. As a foundational proof of concept, this work isolates the phase-retrieval reconstruction step to demonstrate that explicit semantic priors can resolve topological ambiguities where classical methods fail.

1. Introduction

Optical imaging inside and through complex materials remains a significant challenge with broad implications spanning biomedical imaging to autonomous navigation (Gröger et al., 2023). In inhomogeneous media such as biological tissues, microscopic refractive index variations cause strong multiple scattering. This scrambles incident light into a diffused speckle pattern, preventing the formation of standard diffraction-limited images (Katz et al., 2014). Historically, researchers have utilized the optical "memory effect" (ME) for speckle correlations to bypass the scrambling nature of complex media and achieve non-invasive imaging (Freund et al., 1988).

According to the memory effect, when an incident light beam is tilted by a small angle $\Delta\theta$, the transmitted speckle pattern does not decorrelate randomly; rather, it merely translates across the observation plane while maintaining its deterministic structural profile. This high degree of correlation persists as long as the angular extent of the object remains within the medium's isoplanatic patch, typically bounded by $\Delta\theta \ll \lambda/\pi L$, where λ is the illumination wavelength and L is the effective thickness of the scattering layer. Under spatially incoherent illumination, discrete points on a hidden object generate mutually incoherent, shifted copies of the exact same speckle pattern. Because these shifted intensity patterns add linearly without coherent interference, the scattering medium effectively acts as a complex lens.

Consequently, the physical setup functions as a linear, shift-invariant system—meaning the object's image is simply convolved with a fixed pattern. In this case, the system's point-spread function (PSF), or convolutional kernel, is the random speckle pattern itself. By taking the autocorrelation of the captured camera image (see Section 2.1), one effectively computes the convolution of the object's autocorrelation with the speckle's autocorrelation. Because the autocorrelation of a fully developed random speckle pattern is sharply peaked—closely resembling a Dirac delta function superimposed on a uniform background—the autocorrelation of the measured speckle intensity directly isolates the autocorrelation of the hidden object (Katz et al., 2014).

The fundamental challenge of this framework lies in the inherent difficulty of the phase retrieval problem, which is notoriously ill-posed due to the loss of Fourier phase information in intensity-only measurements (Shechtman et al., 2015). Mathematically, recovering a signal from its autocorrelation—defined by the nonlinear operator $A(x) =$

[1]School of Electrical Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, South Korea [2]Department of Bio and Brain Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, South Korea [3]KAIST Institute for Health Science and Technology, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, South Korea. Correspondence to: Mooseok Jang <mooseok@kaist.ac.kr>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

$x \star x$—is complicated by a set of trivial ambiguities: global phase shifts, spatial translations, and conjugate-inversion symmetries. The latter, in particular, often leads to the "twin-image" problem, where traditional solvers produce a superposition of the object and its inverted counterpart (Fienup & Wackerman, 1986).These symmetries create a highly non-convex optimization landscape, causing iterative algorithms like the Gerchberg-Saxton-Fienup (GSF) method (Fienup, 1982)—which recover structural information by alternating between measurement and image domains to enforce physical constraints like non-negativity and data consistency—to frequently stagnate in local minima that bear little to no structural resemblance to the true object.

Furthermore, a significant gap exists between the idealized physical model and real-world experimental conditions. While the theoretical autocorrelation of a random speckle pattern approximates a Dirac delta function (Katz et al., 2014), experimental measurements are plagued by a persistent background DC term, camera shot noise, and deviations from perfect isoplanatism. These factors prevent the measured autocorrelation from being an exact mathematical replica of the object's true autocorrelation. In such degraded settings, unguided reconstruction methods often lack the semantic awareness required to break underlying symmetries or distinguish between distinct classes with similar autocorrelation profiles—such as the digits '6' and '9'. Consequently, traditional solvers fail to bridge the gap between noisy, imperfect physical measurements and the true, contextually accurate object structure.

To overcome the instability of classical optimization, recent research has increasingly explored learning-based methods, offering a structured way to incorporate physical insight directly into computational models. Deep learning paradigms utilize feedforward convolutional neural networks (CNNs) to directly learn the complex mapping from degraded autocorrelations to clean object reconstructions, providing enhanced resilience to noise and experimental artifacts (Li et al., 2018; Metzler et al., 2020).

Recently, diffusion models have emerged as powerful generative priors capable of addressing such ill-posed inverse problems by identifying reconstructions that are both measurement-consistent and aligned with real-world image statistics (Chung et al., 2023; Song et al., 2022). Nonetheless, because diffusion priors rely strictly on image statistics, they can be insufficient in severely ill-posed settings where measurements are highly degraded, often failing to fully break underlying symmetries or guarantee a unique reconstruction. To resolve these remaining ambiguities, additional sources of information are required.

Here, we propose formulating the reconstruction process as a regularized inverse problem solved via a text-guided diffusion model. Let the forward measurement model be $y = A(x) + \epsilon$, where $A(x) = x \star x$ represents the nonlinear autocorrelation operator, x is the true object, y is the observed data and ϵ is the measurement noise. By incorporating descriptive text during the reverse diffusion process, we introduce a semantic constraint that acts as a powerful classification prior. This explicit class awareness steers the sampling trajectory toward solutions that are both measurement-consistent and contextually accurate. Crucially, categorizing the object via text resolves the severe topological ambiguities that cause unguided models to misidentify structures—for instance, easily distinguishing between a '6' and a '9', which possess identical autocorrelation symmetries but belong to distinct semantic classes. This classification mechanism becomes even more essential as object complexity increases. For highly complex objects, a large number of distinct structures can map to nearly identical autocorrelations; thus, **prior knowledge of the target class provides the necessary regularization to prune false structural matches.** Building on recent advances in text-regularized diffusion (TReg) (Kim et al., 2025), we adapt this framework to the physical forward model of speckle-correlation imaging.

Our contributions are as follows:

- **We develop a text-guided diffusion-based framework that actively exploits semantic information to dramatically constrain the solution space of the highly ill-posed inverse problem in imaging through scattering media, thereby enabling reliable recovery of the desired solution.**
- We demonstrate that semantic text guidance successfully breaks the phase-loss symmetries that traditionally plague correlation-based imaging.
- We validate our approach on the MNIST (LeCun et al., 2010) and Quick, Draw! (Jongejan et al., 2017) datasets through **real-world experiments**, outperforming traditional phase-retrieval and unguided deep learning baselines.

2. Methods

This section details our proposed framework for reconstructing hidden objects from scattered light. We divide our methodology into two primary components. First, we establish the physical forward model, utilizing the optical memory effect to frame the complex scattering process as a phase retrieval problem. Second, we introduce our computational inverse solver, which employs a text-guided diffusion model to navigate the highly ill-posed phase-retrieval landscape and resolve inherent structural ambiguities through semantic regularization.

2.1. Physical Forward Model of Speckle-Correlation Imaging

Consider an object $O(\theta)$ hidden behind a scattering medium and illuminated by a spatially incoherent, narrowband light source. A camera captures the resulting scattered intensity $I(\theta)$. Due to the optical memory effect, as long as the object is confined within the angular range $\Delta\theta \ll \lambda/\pi L$ (where L is the medium thickness and λ is the wavelength), the scattering process operates entirely as a shift-invariant convolution. Under these conditions, the captured camera image is a simple convolution of the object with the system's speckle point-spread function (PSF) $S(\theta)$,

$$I(\theta) = [O * S](\theta) \tag{1}$$

To extract structural information, we calculate the spatial autocorrelation ($\star$) of the recorded image:

$$[I \star I](\theta) = [(O \star O) * (S \star S)](\theta) \tag{2}$$

Crucially, the spatial autocorrelation of a fully developed random speckle pattern mathematically approximates a Dirac delta function $\delta(\theta)$ superimposed on a uniform background constant C:

$$[S \star S](\theta) \approx \delta(\theta) + C \tag{3}$$

Because convolving any function with a Dirac delta returns the original function, applying this property reveals that the measured image autocorrelation is effectively identical to the autocorrelation of the hidden object:

$$[I \star I](\theta) \approx [O \star O](\theta) + C \tag{4}$$

Thus, our fundamental inverse problem is established: we seek to recover the true object state $x = O(\theta)$ from the physically constrained measurement $y = [I * I](\theta)$. According to the Wiener-Khinchin theorem (Khintchine, 1934), the Fourier transform of this spatial autocorrelation yields the power spectrum of the hidden object, $|\mathcal{F}\{O(\theta)\}|^2$. Because this relation provides only the magnitude of the object's frequency components while completely discarding the Fourier phase, recovering the original structure formally reduces to a classical phase retrieval problem.

2.2. Text-Guided Diffusion model for Solving Inverse Problem

2.2.1. Preliminaries

Directly performing diffusion in high-dimensional pixel space is computationally prohibitive for real-time reconstruction. To mitigate this, our framework utilizes a Latent Diffusion Model (LDM) that operates in a lower-dimensional manifold z, where $\dim(z) \ll \dim(x)$. This perceptual compression relies on a variational autoencoder (VAE) consisting of an encoder $\mathcal{E}$ that maps the input image $x \in \mathbb{R}^{H\times W\times 3}$ to a latent representation $z = \mathcal{E}(x) \in \mathbb{R}^{h\times w\times c}$, and a decoder $\mathcal{D}$ that learns to map these latents back to the pixel domain such that $\mathcal{D}(\mathcal{E}(x)) \approx x$. In this compressed space, a forward diffusion process q systematically perturbs the clean latent z_0 into Gaussian noise z_T over T timesteps:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t} z_0, (1-\bar{\alpha}_t)I) \tag{5}$$

A time-conditional U-Net denoiser $\epsilon_\theta(z_t, t, c)$ is then trained to predict the added noise given a text embedding condition c. By minimizing the epsilon-matching objective, the model learns the score of the latent data distribution.

Building on recent advances in text-regularized diffusion (TReg (Kim et al., 2025)), the reconstruction process can be formulated as a proximal optimization problem. The unified objective function jointly optimizes the pixel-space image and its latent representation z:

$$\min_{x,z} \mathcal{L}_{MAP}(x,z) + \gamma \mathcal{L}_{TReg}(z) \quad \text{s.t.} \quad x = \mathcal{D}_\varphi(z) \tag{6}$$

The first term ensures data consistency in the pixel domain and adherence to the pre-trained latent manifold:

$$\mathcal{L}_{MAP}(x,z) = \frac{\|y - A(x)\|^2}{2\sigma^2} + \frac{\|z - \mathcal{E}_\phi(x)\|^2}{2\sigma_z^2} \tag{7}$$

The second term provides crucial semantic regularization, actively steering the latent state toward structures that match the linguistic description:

$$\mathcal{L}_{TReg}(z) = \|z - \hat{z}_{0|t}(c)\|^2 \tag{8}$$

Here, $\hat{z}_{0|t}(c)$ serves as a text-conditioned pivot representing the clean latent estimate at timestep t, computed using Tweedie's formula:

$$\hat{z}_{0|t}(c) = \frac{z_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(z_t, c, t)}{\sqrt{\bar{\alpha}_t}} \tag{9}$$

2.2.2. Text-Guided Diffusion for Speckle-Correlation Imaging

As simultaneously optimizing both variables is highly complex, we decouple the process using an alternating variable splitting approach. This anchors the solution space via the text-guided prior in the latent domain, decodes and optimizes in the pixel domain to satisfy the physical measurement, and merges the physics-constrained result with the semantic anchor. To balance computational efficiency with the stability of the generative prior, these heavy data consistency and semantic refinement steps are executed on a periodic schedule (e.g., every third denoising timestep) rather than at every step.

During an active refinement step at timestep t, the inference sequence proceeds through the following structured operations:

- **Text-Guided Tweedie Estimation:** The text prompt c is injected into the UNet denoiser using Classifier-Free Guidance (CFG), steering the noise prediction toward the semantic concept. Applying Tweedie's formula yields the text-conditioned clean latent estimate $\hat{z}_{0|t}(c)$, restricting the structural solution space (9).

- **Grayscale Compression:** Because phase retrieval is highly unstable and prone to chromatic artifacts, tensors in the image domain are compressed into a grayscale representation. To preserve dimensional compatibility with the 3-channel architecture, the tensor is averaged across its color channels to compute a single-channel mean, which is then replicated across the RGB channels:

$$x^{gray} = [\bar{x}, \bar{x}, \bar{x}] \quad \text{where} \quad \bar{x} = \frac{1}{3} \sum_{i \in \{R,G,B\}} x^{(i)} \tag{10}$$

  For latent vectors, we apply an identical operation by first decoding the latent into pixel space, computing the channel-wise average, and then replicating this mean before encoding it back into the latent dimension:

$$\begin{aligned} \bar{z} &= \frac{1}{3} \sum_{i \in \{R,G,B\}} (\mathcal{D}_\varphi(z))^{(i)} \\ z^{gray} &= \mathcal{E}_\varphi\left([\bar{z}, \bar{z}, \bar{z}]\right) \end{aligned} \tag{11}$$

- **Data-Consistency Optimization:** To isolate structural optimization from color generation, data consistency is strictly enforced in the pixel domain. Specifically, the grayscale pixel-space estimate $\hat{x}_{0|t}^{gray}$ is detached from the diffusion computational graph and set as the active optimization variable. The Adam optimizer calculates the gradient of the composite loss strictly with respect to $\hat{x}_{0|t}^{gray}$ (i.e., $\nabla_{\hat{x}_{0|t}^{gray}} \mathcal{L}_{DC}$) and iteratively updates it to minimize:

$$\mathcal{L}_{DC} = \text{MSE}(y, A(\hat{x}_{0|t}^{gray})) + \alpha \text{TV}(\hat{x}_{0|t}^{gray}) \tag{12}$$

  where y is the experimental spatial autocorrelation, A is the nonlinear autocorrelation operator ($A(x) = x \star x$), and TV applies Total Variation regularization to suppress high-frequency background noise. Once optimized, this updated pixel-space image is re-encoded back into the latent space: $\hat{z}_0(y) = \mathcal{E}_\phi(\hat{x}_0(y))$.

- **Adaptive Negation:** To prevent text descriptions from introducing semantic noise, the null-text embedding $c_\emptyset$, which represents the concepts to be suppressed, is dynamically updated. Using a pre-trained CLIP image encoder $\mathcal{T}_{img}$, we minimize the cosine similarity (S_C) between the current image representation and the null embedding via gradient descent:

$$c_\emptyset \leftarrow c_\emptyset - \eta \nabla_{c_\emptyset} S_C(\mathcal{T}_{img}(\hat{x}_0(y)), c_\emptyset) \tag{13}$$

  This continuous negation captures invalid concepts and artifacts within $c_\emptyset$. Crucially, this updated null-embedding is subsequently injected into the Classifier-Free Guidance (CFG) formulation:

$$\hat{\epsilon}_\theta = \epsilon_\theta(z_t, c_\emptyset, t) + \omega(\epsilon_\theta(z_t, c, t) - \epsilon_\theta(z_t, c_\emptyset, t)) \tag{14}$$

  By actively pushing the sampling trajectory away from the dynamically updated $c_\emptyset$, the framework ensures the text guidance tightly follows the evolving physical structure. During the inference stage, this CFG-modified noise prediction $\hat{\epsilon}_\theta$ serves two critical functions: it is immediately utilized in Tweedie's formula (Equation 9) to generate the text-conditioned structural pivot $\hat{z}_{0|t}(c)$, and it dictates the noise required to step backward in time to $t-1$ (Equation 16).

- **Latent Update and Interpolation:** To maintain the stability of the generative prior, the final clean latent estimate is calculated as an interpolation between this measurement-consistent latent and the original text-guided pivot:

$$\hat{z}_0^{ema} = \bar{\alpha}_{t-1} \hat{z}_0(y) + (1 - \bar{\alpha}_{t-1}) \hat{z}_{0|t}(c) \tag{15}$$

  Finally, to proceed to the next reverse sampling step, this merged clean estimate is re-noised to form the latent state at timestep $t-1$. The total noise $\tilde{\epsilon}_t$ is added back to $\hat{z}_0^{ema}$ to compute z_{t-1}:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{z}_0^{ema} + \sqrt{1 - \bar{\alpha}_{t-1}} \tilde{\epsilon}_t \tag{16}$$

  Here, $\tilde{\epsilon}_t$ denotes the total noise given by:

$$\tilde{\epsilon}_t := \sqrt{1 - \bar{\alpha}_{t-1}} \hat{\epsilon}_\theta + \sqrt{\bar{\alpha}_{t-1}} \epsilon \tag{17}$$

  where $\epsilon \sim \mathcal{N}(0, I)$ is standard Gaussian noise, and $\hat{\epsilon}_\theta$ is the CFG-modified noise prediction. This updated noisy latent z_{t-1} is then fed into the next iteration of the diffusion sampling loop, guaranteeing convergence to a physics-constrained reconstruction.

By systematically alternating between text-guided latent prediction and physics-constrained pixel optimization, the sampling trajectory strictly converges to a reconstruction that effectively breaks phase-loss symmetries while maintaining high fidelity to the physical speckle measurement.

Algorithm 1 Text-Guided Diffusion for Speckle-Correlation Imaging

Require: Pretrained LDM ϵ_θ, Null text embedding $c_\emptyset$, Text embedding c, VAE encoder $\mathcal{E}_\phi$, VAE decoder $\mathcal{D}_\varphi$, Forward operation A, Measurement y, CLIP image encoder $\mathcal{T}_{img}$, Active refinement schedule Γ
Initialization $z_T \sim \mathcal{N}(0, \mathbf{I})$
for $t \in [T, 1]$ **do**
  $\hat{\epsilon}_\theta \leftarrow \epsilon_\theta(z_t, c_\emptyset, t) + \omega\left(\epsilon_\theta(z_t, c, t) - \epsilon_\theta(z_t, c_\emptyset, t)\right)$
  $\hat{z}_{0|t}(c) \leftarrow (z_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_\theta)/\sqrt{\bar{\alpha}_t}$
  $\tilde{\epsilon}_t \leftarrow$ Compute noise using (17)
  if $t \in \Gamma$ **then**
    // Grayscale Compression & Data-Consistency
    $\bar{z} \leftarrow \frac{1}{3}\sum_{i\in\{R,G,B\}}(\mathcal{D}_\varphi(\hat{z}_{0|t}(c)))^{(i)}$
    $\hat{x}_{0|t}^{gray} \leftarrow [\bar{z}, \bar{z}, \bar{z}]$
    $\hat{x}_0(y) \leftarrow$ Optimize Equation (12) w.r.t. $\hat{x}_{0|t}^{gray}$
    $\hat{z}_0(y) \leftarrow \mathcal{E}_\phi(\hat{x}_0(y))$
    $\hat{z}_0^{ema} \leftarrow \bar{\alpha}_{t-1}\hat{z}_0(y) + (1-\bar{\alpha}_{t-1})\hat{z}_{0|t}(c)$
    $z_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\hat{z}_0^{ema} + \sqrt{1-\bar{\alpha}_{t-1}}\tilde{\epsilon}_t$
    // Adaptive Negation
    $c_\emptyset \leftarrow c_\emptyset - \eta\nabla_{c_\emptyset} S_C(\mathcal{T}_{img}(\hat{x}_0(y)), c_\emptyset)$
  else
    $z_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\hat{z}_{0|t}(c) + \sqrt{1-\bar{\alpha}_{t-1}}\tilde{\epsilon}_t$
  end if
end for

3. Experimental Setup and Implementation Details

3.1. Optical Setup

To physically validate our approach, we constructed a speckle-correlation imaging system, as illustrated in Figure 1. A laser beam is first passed through a rotating diffuser to synthesize spatially incoherent illumination. This incoherent light then illuminates the target object, which is displayed on a Spatial Light Modulator (SLM). The light propagating from the object directly strikes a visually opaque 220-grit ground-glass diffuser, which acts as the complex scattering medium. As the incident wavefront traverses this highly inhomogeneous material, it experiences severe multiple scattering, completely scrambling the object's spatial information into a complex, seemingly random speckle field. Finally, a camera positioned behind the diffuser records the resulting scrambled speckle intensity pattern. As demonstrated in Figure 1, while the raw camera image lacks any visual spatial information, its computed spatial autocorrelation closely matches the ideal object autocorrelation. This experimentally obtained speckle autocorrelation serves directly as the measurement y in our data consistency term.

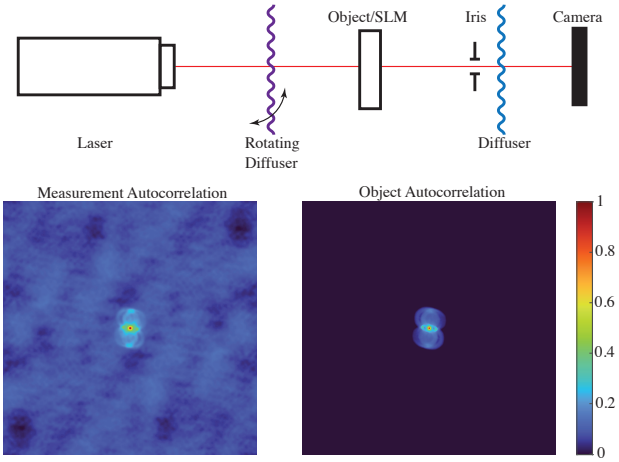

Figure 1. Experimental Validation Setup. The optical configuration consists of the laser, scattering medium, and camera used to capture the scrambled speckle intensity measurements.

3.2. Data Preprocessing and Validation

Before passing the experimentally captured speckle intensity patterns to our diffusion-based inverse solver, the raw camera images undergo a computational preprocessing pipeline to mitigate experimental artifacts. First, a dark-frame offset image is subtracted from the raw speckle measurements to remove ambient light and camera sensor dark current, with negative values clipped to zero. The images are then cropped to a consistent 512×512 pixel region of interest. To correct for the slowly varying envelope (or halo) of the scattered light pattern, the image is spatially normalized by dividing it by a low-pass-filtered version of itself (utilizing a Gaussian kernel). Finally, the normalized image is smoothed with a narrow Gaussian kernel ($\sigma = 0.5 - 2$ pixels) to suppress high-frequency noise before computing the spatial autocorrelation.

To validate our optical setup and ensure the targets fall well within the optical memory effect range, we quantitatively compared the autocorrelation of the preprocessed speckle patterns against the theoretical autocorrelation of the ground-truth objects displayed on the SLM. As visually demonstrated in Figure 1, the autocorrelations exhibit a strong structural alignment, achieving a matching score of 0.65 Pearson correlation coefficient on average.

3.3. Datasets

We fine-tuned and evaluated our proposed method on two standard datasets to simulate complex structural recovery. For both datasets, the text conditioning prompts were uniformly formatted as "image of [label]" (e.g., "image of 7"

or "image of apple"). For training, we sampled 800 images per class for each dataset, and for testing, we sampled 100 images per class for each dataset.

- **MNIST:** We used the 10 standard classes of handwritten digits (0–9). To match the spatial dimensions of our forward model, the original 28×28 pixel images were resized to 56×56 and subsequently zero-padded to 512×512 pixels.

- **Quick, Draw!:** To test the limits of our phase-retrieval framework on highly complex, continuous shapes, we selected a diverse subset of 10 sketch classes (apple, banana, cloud, envelope, eyeglasses, hand, paper clip, rainbow, smiley face, and zigzag). The original 256×256 pixel sketches, rendered with a stroke width of 18, were downsampled to 50×50 pixels and zero-padded to 512×512 pixels.

3.4. Training Details

For the diffusion prior, we utilized the architecture provided by the TReg framework, which uses the Stable Diffusion v1.5 pipeline (Rombach et al., 2022). Specifically, we first fine-tuned the Variational Autoencoder (VAE) to capture domain-specific latent representations, and subsequently fine-tuned the U-Net utilizing this updated VAE. The raw speckle patterns were experimentally captured using our optical setup and preprocessed following Section 3.2, and the measurement y was obtained by computing the spatial autocorrelation of these physical speckle images.

To adapt the generative prior to our specific target domains (MNIST and Quick, Draw!), we employed an identical two-stage fine-tuning strategy for both datasets. First, the VAE was initialized from the `stabilityai/sd-vae-ft-mse` checkpoint and fine-tuned for 50 epochs. We utilized a learning rate of 4.5×10^{-6} paired with a cosine learning rate scheduler, 2 gradient accumulation steps, and `bf16` mixed precision optimization.

Following the VAE refinement, the U-Net was initialized from the `runwayml/stable-diffusion-v1-5` baseline and fine-tuned on 512×512 resolution images for a maximum of 15,000 steps. The U-Net training was conducted with a constant learning rate of 1×10^{-5}, a per-device batch size of 1, and 4 gradient accumulation steps (yielding an effective batch size of 4). To optimize computational overhead and memory usage during this phase, both gradient checkpointing and memory-efficient attention (via `xformers`) were enabled.

4. Results and Discussion

4.1. Reconstruction on MNIST and Quick, Draw!

Figure 2 visualizes the qualitative results on the MNIST dataset. The measurements y contain no discernible spatial information. Our text-guided diffusion model accurately reconstructs the digits, preserving both the topological structure and stroke thickness. To further test the algorithm, we evaluated reconstructions on the Quick, Draw! dataset, which features continuous, non-sparse strokes that heavily violate the sparsity priors typically required by phase-retrieval techniques. Figure 3 demonstrates that the text-guided diffusion framework successfully restores the semantic shape of the sketches on these continuous objects.

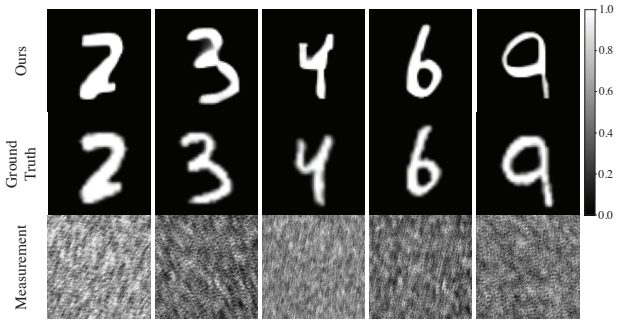

Figure 2. Qualitative reconstruction results on the MNIST dataset. Columns from top to bottom: Our Proposed Method (Text-Guided Diffusion), Ground Truth, Measurement (Scattered Speckle Patterns).

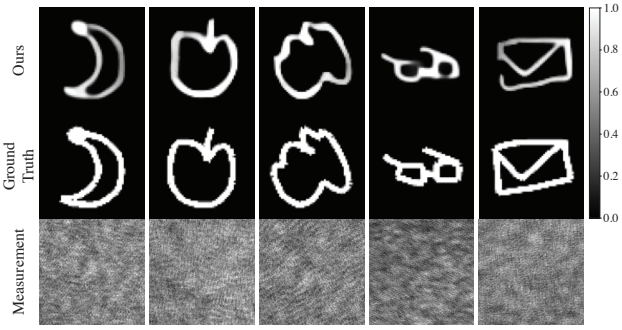

Figure 3. Qualitative reconstruction results on the Quick, Draw! dataset, highlighting the robustness of our method on complex drawings. Columns from top to bottom: Our Proposed Method (Text-Guided Diffusion), Ground Truth, Measurement (Scattered Speckle Patterns).

4.2. Related Work and Baseline Comparison

To contextualize our contributions and establish our experimental baselines, we review the two primary paradigms for correlation imaging through scattering media: classical phase retrieval and feedforward deep learning.

Speckle-Correlation Imaging (Classical Baseline). Non-invasive single-shot imaging through scattering layers via speckle correlations was pioneered by Katz et al. (2014). By exploiting the optical memory effect, they demonstrated that the autocorrelation of a speckle pattern essentially matches the hidden object's autocorrelation, allowing reconstruction via the Fienup (GSF) phase-retrieval algorithm. However, this approach relies heavily on classical optimization, which frequently stagnates and fails for dense, complex, or non-sparse objects, particularly in low signal-to-noise ratio (SNR) conditions. In our quantitative evaluations, the method of Katz et al. represents the classical physics-based baseline.

Deep Learning for Phase Retrieval (State-of-the-Art Baseline). To overcome the limitations of traditional phase retrieval in photon-starved environments, Metzler et al. (2020) introduced Deep-inverse correlography. While their system relies on NLoS correlography under coherent illumination—distinct from our spatially incoherent correlation imaging—their learned reconstruction serves as a highly relevant state-of-the-art baseline for solving the degraded phase retrieval problem. To train their convolutional neural network (U-Net), Metzler et al. (2020) used spectral estimation theory to derive a rigorous noise model that explicitly accounts for finite-sample approximation error, photon shot noise, ambient illumination, and finite apertures. They then sampled from this physical noise model to generate synthetic training data. While highly robust to noise compared to classical methods, their approach utilizes a purely feedforward network trained primarily on sparse, unstructured edge data. Consequently, it lacks the ability to disambiguate highly complex or dense structures. This serves as our state-of-the-art deep learning baseline.

Our Advantage. In contrast to these existing methods, our text-guided diffusion framework leverages strong, semantic generative priors combined with explicit data-consistency optimization. Due to the sparsity of the target objects, pixel-wise metrics such as MSE and PSNR do not adequately reflect overall visual fidelity, whereas LPIPS offers a more realistic, perceptually grounded evaluation. As demonstrated in our quantitative results (Section 4.1), our method achieves state-of-the-art reconstruction LPIPS of **0.2338** on the MNIST dataset, surpassing the Katz et al. (2014) baseline of **0.5650** LPIPS and consistently outperforming the Metzler et al. (2020) approach of **0.4970** LPIPS. The semantic text guidance prevents the structural stagnation of classical phase retrieval and provides a substantially stronger, context-aware regularizing prior than unstructured deep learning methods.

Table 1. Quantitative reconstruction evaluation on MNIST and Quick, Draw! datasets. **Bold** indicates the best performance, and <u>underline</u> indicates the second best.

Method	PSNR ↑	LPIPS ↓	MSE ↓
MNIST Dataset			
Katz et al.	12.0754	0.5650	0.0659
Metzler et al.	<u>15.3760</u>	<u>0.4970</u>	<u>0.0329</u>
Ours	**16.2214**	**0.2338**	**0.0263**
Quick, Draw! Dataset			
Katz et al.	9.1260	<u>0.5622</u>	0.1270
Metzler et al.	<u>10.1816</u>	0.5938	<u>0.0992</u>
Ours	**11.3629**	**0.2730**	**0.0813**

While our experimental validations demonstrate high fidelity, it is crucial to acknowledge that the idealized forward imaging model—which assumes perfect shift-invariant convolution—is rarely perfectly matched in practice. Inevitable deviations arise from real-world experimental conditions, including camera shot and read noise, finite numerical aperture limitations, and slight breakdowns in perfect isoplanatism across the scattering diffuser. Consequently, the experimental speckle autocorrelation is never a mathematically perfect copy of the ground-truth autocorrelation. However, these physical discrepancies highlight the exact strength of our proposed framework. When the strict assumptions of the forward imaging model degrade, classical phase-retrieval algorithms inevitably stagnate. In contrast, our text-guided diffusion approach seamlessly resolves these physical ambiguities. By providing a robust semantic prior, the text conditioning supplies the critical contextual information required to accurately steer the reverse sampling trajectory, effectively bridging the gap between noisy, imperfect physical measurements and the true object structure. As illustrated in Figure 4, our proposed method achieves substantial qualitative improvements over the compared works.

4.3. Ablation Study on Text Guidance (CFG)

To isolate the impact of textual semantic regularization on resolving phase ambiguity, we performed an ablation study modulating the classifier-free guidance (CFG) scale. Because the fine-tuned diffusion backbone inherently possesses a highly expressive image prior, this ablation is critical to underscore that text conditioning is not merely an auxiliary feature, but the primary mechanism driving successful and unique structural reconstruction. As shown in Figure 5, without text guidance (CFG = 0), the diffusion model relies solely on this image prior and the autocorrelation constraint. Although unguided reconstruction may occasionally yield plausible results, it lacks robustness and is highly sus-

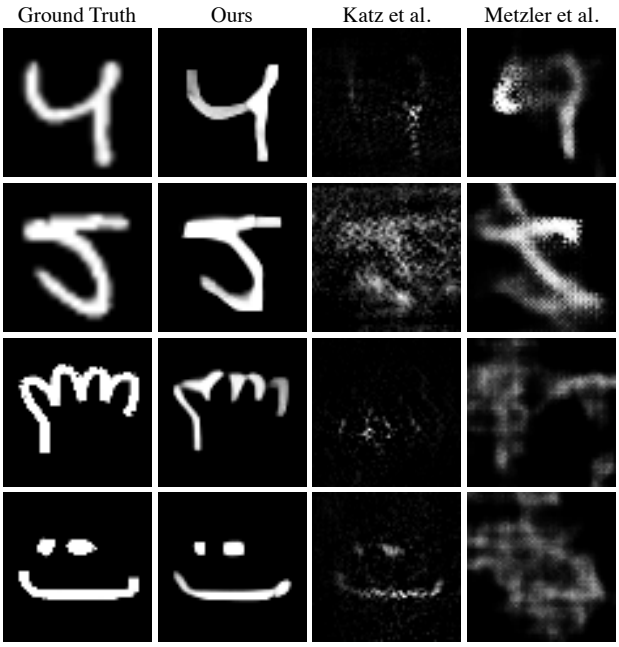

Figure 4. Quantitative comparison against (Katz et al., 2014) and (Metzler et al., 2020).

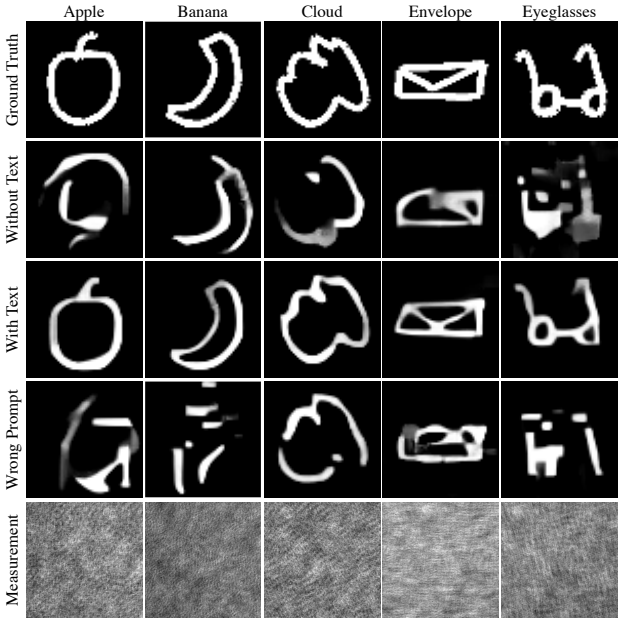

Figure 5. Ablation study demonstrating the effect of Classifier-Free Guidance (CFG) scale on reconstruction fidelity.

ceptible to topological ambiguities; however, as the CFG scale increases, the text prompt aggressively prunes the solution space, yielding consistently higher reconstruction fidelity and a strong structural alignment with the ground truth. Furthermore, we evaluated our framework's robustness to incorrect text conditioning by deliberately applying mismatched prompts: we used 'image of zigzag' for non-zigzag objects, and 'image of apple' for zigzag objects. The network's subsequent struggle with reconstruction (Figure 5) underscores the critical importance of text conditioning in this task.

Table 2. Ablation study on Quick, Draw! dataset. **Bold** indicates the best performance, and underline indicates the second best.

Method	PSNR ↑	LPIPS ↓	MSE ↓
Quick, Draw! Dataset			
Wrong Prompt	9.8646	0.3686	0.1090
Without Text	<u>10.1351</u>	<u>0.3483</u>	<u>0.1034</u>
With Text	**11.3629**	**0.2730**	**0.0813**

5. Conclusion

We have presented a novel text-guided diffusion framework for non-invasive optical imaging through scattering media. By modeling speckle-correlation imaging as an autocorrelation inverse problem, we leveraged the generative power of diffusion models and the semantic constraint of natural language to break phase ambiguities. Our experimental validation on the MNIST and Quick, Draw! datasets confirms that this approach significantly outperforms classical phase-retrieval algorithms, particularly for complex objects. This work establishes a robust pathway toward deep-tissue biomedical imaging and computational microscopy.

Moving forward, our future research will focus on two critical directions: extending this framework to operate beyond the optical memory effect (overcoming the highly restrictive isoplanatic patch) and scaling the method to reconstruct highly dense, non-sparse objects (such as continuous grayscale natural scenes and complex biological tissues). By explicitly targeting these out-of-memory-effect and high-density challenges with robust semantic diffusion priors, this work lays a pathway toward calibration-free foundational methods for real-world deep-tissue biomedical imaging and advanced computational microscopy.

Acknowledgments

This research was supported by the National Research Foundation of Korea grant funded by the Korean government (MSIT) (Grant Nos. RS-2026-25479811).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and Computational Optics. There are many potential societal consequences of our work, particularly in enabling non-invasive biomedical diagnostics and deep-tissue imaging, none of which we feel must be specifically highlighted here as a negative ethical concern.

References

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

Fienup, J. R. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21(15):2758–2769, Aug 1982.

Fienup, J. R. and Wackerman, C. C. Phase-retrieval stagnation problems and solutions. *J. Opt. Soc. Am. A*, 3(11): 1897–1907, Nov 1986.

Freund, I., Rosenbluh, M., and Feng, S. Memory effects in propagation of optical waves through disordered media. *Phys. Rev. Lett.*, 61:2328–2331, Nov 1988.

Gröger, A., Pedrini, G., Claus, D., Alekseenko, I., Gloeckler, F., and Reichelt, S. Advantages of holographic imaging through fog. *Appl. Opt.*, 62(10):D68–D76, Apr 2023.

Jongejan, J., Rowley, H., Kawashima, T., Kim, J., and Fox-Gieg, N. The Quick, Draw! Dataset, 2017. URL https://github.com/googlecreativelab/quickdraw-dataset.

Katz, O., Heidmann, P., Fink, M., and Gigan, S. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature Photonics*, 8(10):784–790, Oct 2014.

Khintchine, A. Korrelationstheorie der stationären stochastischen prozesse. *Mathematische Annalen*, 109 (1):604–615, Dec 1934.

Kim, J., Park, G. Y., Chung, H., and Ye, J. C. Regularization by texts for latent diffusion inverse solvers. In *The Thirteenth International Conference on Learning Representations*, 2025.

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Li, Y., Xue, Y., and Tian, L. Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media. *Optica*, 5(10):1181–1190, Oct 2018.

Metzler, C. A., Heide, F., Rangarajan, P., Balaji, M. M., Viswanath, A., Veeraraghavan, A., and Baraniuk, R. G. Deep-inverse correlography: towards real-time high-resolution non-line-of-sight imaging. *Optica*, 7(1):63–71, Jan 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. doi: 10.1109/MSP.2014.2352673.

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022.