GenPIE: A Time-Resolved Plenoptic Imager

ZIHENG WANG, The Chinese University of Hong Kong, Shenzhen, China
SIYUAN SHEN, ShanghaiTech University, China
HUANYU XU, ShanghaiTech University, China
KAICHUN QIAO, ShanghaiTech University and Deemos Technology, China
LONGWEN ZHANG, ShanghaiTech University and Deemos Technology, China
QIXUAN ZHANG, ShanghaiTech University and Deemos Technology, China
QILIN SUN, The Chinese University of Hong Kong, Shenzhen and Point Spread Technology, China
SHIYING LI*, ShanghaiTech University, China
JINGYI YU*, ShanghaiTech University, China

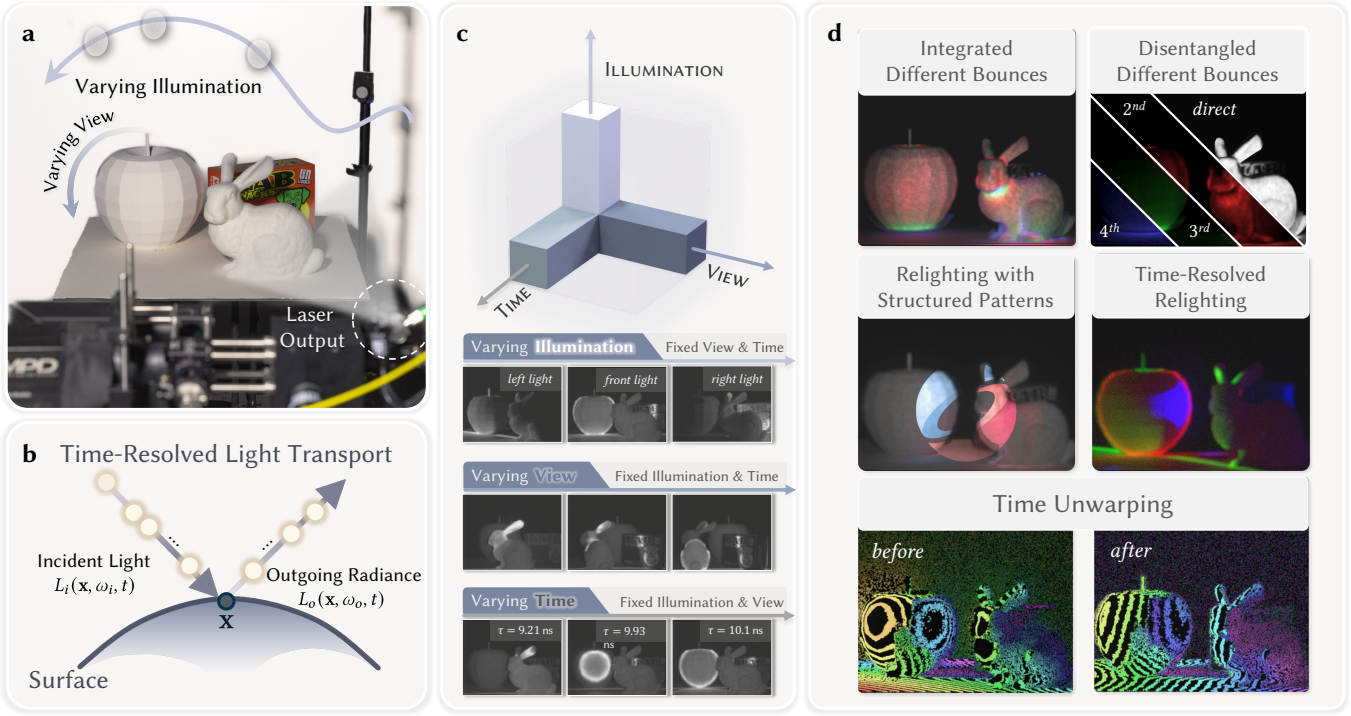

Fig. 1. *GenPIE* captures and recovers plenoptic light transport. **(a)** We build a time-resolved imaging system to capture plenoptic light transport from configurable views and illuminations. **(b)** The underlying physical light transport is described by the Transient Rendering Equation. **(c)** By independently varying viewing angles and illuminations, and recording temporal dimension of light transport, we capture slices of the high-dimensional plenoptic function. **(d)** We demonstrate physically grounded applications, including disentangling multi-bounce light transport, time-resolved relighting, and time unwarping.

Capturing the full plenoptic light transport across spatial, angular, and temporal dimensions has long been a pursuit in computational imaging, yet it remains fundamentally constrained by the high dimensionality of the sampling space and the physical inaccessibility of scene regions due to self-occlusions. While time-resolved imaging records the temporal axis, existing methods are bottlenecked by the combinatorial complexity of the plenoptic function. This high dimensionality makes dense omni-dimensional sampling physically prohibitive. Simultaneously, tight coupling between illumination and viewpoint in current systems also precludes the full acquisition of plenoptic light transport. In this work, we present GenPIE, a Generative Plenoptic Imager designed to bridge the gap between sparse physical observations and high-dimensional light transport. We introduce a decoupled laser-detector hardware setup that enables independent control over illumination and detection, allowing for active probing of indirect light

*Corresponding authors.

Authors' Contact Information: Ziheng Wang, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, zihengwang3@link.cuhk.edu.cn; Siyuan Shen, ShanghaiTech University, Shanghai, China; Huanyu Xu, ShanghaiTech University, Shanghai, China; Kaichun Qiao, ShanghaiTech University and Deemos Technology, Shanghai, China; Longwen Zhang, ShanghaiTech University and Deemos Technology, Shanghai, China; Qixuan Zhang, ShanghaiTech University and Deemos Technology, Shanghai, China; Qilin Sun, The Chinese University of Hong Kong, Shenzhen and Point Spread Technology, Shenzhen, China; Shiying Li, ShanghaiTech University, Shanghai, China, lishy1@shanghaitech.edu.cn; Jingyi Yu, ShanghaiTech University, Shanghai, China, yujingyi@shanghaitech.edu.cn.

paths. To overcome the ill-posedness of sparse sampling and physical blind spots, we propose a generative inverse transient rendering framework. Our approach leverages 3D foundation models to provide strong semantic and 3D geometric priors for initialization, which are subsequently refined through a differentiable transient path tracer to ensure physically grounded adherence to the Transient Rendering Equation. We demonstrate that GenPIE supports a range of applications that are challenging for steady-state or purely neural methods, including disentangling multi-bounce light transport directly from captured transient videos, time unwarping, and time-resolved relighting. The project page is at https://wangzh1.github.io/GenPIE.

CCS Concepts: • **Computing methodologies → Computational photography**.

Additional Key Words and Phrases: Computational Imaging, Transient Imaging, Time-Resolved Light Transport

1 INTRODUCTION

An image serves as both a record of the objective physical realm and a reflection of our subjective inner world. The mechanics of how an image is acquired are strictly physical, governed by the Rendering Equation [Kajiya 1986], which dictates how light interacts with matter and applies equally to both the biological eye and the digital sensor. Yet, an image extends beyond physics; humans possess an inherent cognitive capacity to infer the unseen. We instinctively reconstruct missing information, such as occluded objects, novel viewpoints, and unobserved lighting conditions, by drawing upon learned experience. This cognitive aspect is vividly manifested in recent generative image synthesis [Gao et al. 2025; OpenAI 2025; Wu et al. 2025], where models construct pixels to fill semantic gaps using features distilled from vast datasets. However, such generative processes prioritize visual fidelity over physical correctness, mimicking the result without adhering to the underlying process.

Converging these two distinct perspectives leads to what we term a Physically Grounded Generative Imager (PGGI). On the physical front, PGGI aims to reproduce the intricate interactions between light and scene, strictly adhering to the unbending laws of light transport. Simultaneously, PGGI is generative, leveraging the semantic intuition of pre-trained foundation models to recover transport information in regions that are physically unobservable or ill-posed for traditional reconstruction. The power of combining sparse physical sensing with learned priors is exemplified by the recent "minimalist vision" framework [Klotz and Nayar 2024], which demonstrates that an incredibly small number of pixels (e.g., just eight) coupled with task-specific inference can achieve robust vision tasks. We extend this philosophy from 2D recognition to the complex domain of dense 3D light transport.

In this work, we apply the PGGI paradigm to plenoptic imaging—resolving the complete light transport process across spatial, angular, and temporal dimensions. The pursuit of plenoptic imaging has a rich history, evolving from early techniques in Image-Based Rendering (IBR) [McMillan and Bishop 1995], light field cameras [Levoy and Hanrahan 1996], and light stage systems [Debevec et al. 2000], which focused primarily on spatial and angular diversity. However, the field eventually realized that to fully characterize complex transport such as subsurface scattering and inter-reflections, it is essential to record the time dimension. This capability was enabled by the advent of ultrafast detectors, such as Single-Photon Avalanche Diodes (SPADs) and streak cameras, which capture picosecond-scale Time-of-Flight (ToF) information. Since the seminal single-view visualization by [Velten et al. 2013], research has expanded toward multi-view transient modeling and video generation [Luo et al. 2025; Malik et al. 2025, 2024, 2023a]. In parallel, analyzing third-bounce (or higher) light transport has further unlocked the ability to image Non-Line-of-Sight (NLOS) scenes [Buttafava et al. 2015; Lindell et al. 2019; Liu et al. 2020; Royo et al. 2023b; Shen et al. 2021; Velten et al. 2012; Xin et al. 2019].

Despite these efforts, acquiring the complete plenoptic function remains hindered by three fundamental challenges. First, selective sampling leads to information loss and entanglement. For instance, single-view methods [Velten et al. 2013] suffer from occlusion, while recent multi-view approaches [Malik et al. 2025, 2024] tightly couple illumination and viewpoint, preventing effective disentanglement of scene attributes. Second, complex geometry and self-occlusion render parts of the plenoptic domain physically inaccessible [Shen et al. 2025], making standard acquisition methods incapable of observing the full range of transport phenomena. Finally, dense omni-dimensional sampling creates a prohibitive system burden, where the resulting data explosion overwhelms current storage and computational resources.

To address these challenges, we employ a decoupled transient acquisition system that enables independent control over illumination and detection. The system supports two configurations: a confocal imaging mode for acquiring initial scene geometry via LiDAR-like scanning, and a decoupled imaging mode that separates illumination from detection to actively probe broader indirect transport paths. This flexibility allows for adaptive sampling across space, time, viewpoint, and lighting, tailored to scene complexity. However, even this enhanced flexibility cannot fully circumvent the combinatorial explosion of high-dimensional observation or the blind spots caused by physical self-occlusion; consequently, practical acquisition inevitably remains sparse and incomplete.

To bridge the gap between sparse observation and dense reconstruction, we introduce a generative inverse transient rendering framework that integrates learned priors with physically grounded light transport modeling. Specifically, we leverage a 3D Foundation Model to infer high-quality geometric initialization from sparse measurements, effectively generating plausible structures for unobservable regions. To ensure fidelity to the actual measurements, we refine this reconstruction using a differentiable path tracing renderer that explicitly simulates time-resolved light propagation governed by the Transient Rendering Equation [Smith et al. 2008]. Additionally, we incorporate a neural field–based compensation module to implicitly model system-specific effects, such as the instrument response function.

This culminates in the development of *GenPIE* (Generative Plenoptic Imager), the first system capable of reproducing the complete light transport of real-world scenes exhibiting diverse materials and varying geometric complexities. Comprehensive experiments demonstrate that GenPIE captures transient data rich enough to support reliable plenoptic reconstruction. Specifically, the synergy of foundation model-generated priors with precise physical modeling effectively recovers occluded scene information and disentangles

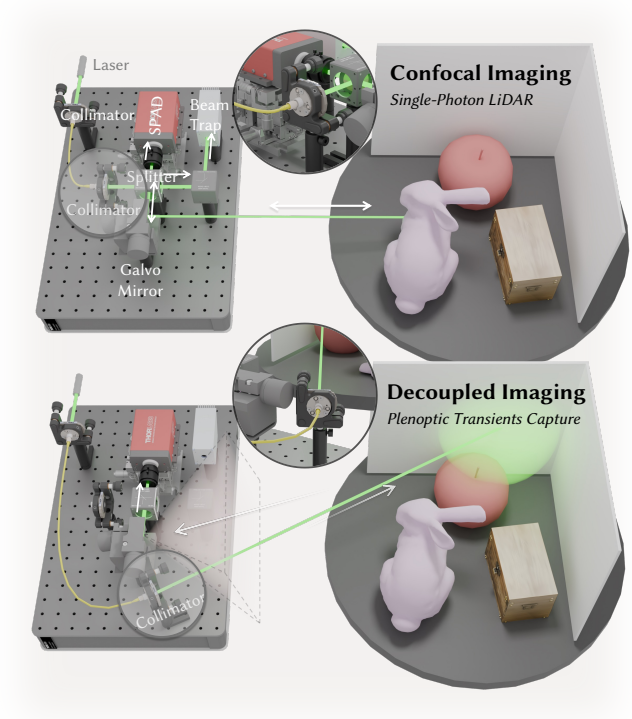

Fig. 2. Illustration of our dual-mode transient imaging system. **Top** (confocal imaging): The illumination and detection paths are optically co-aligned via a beam splitter. This mode captures direct time-of-flight signals that provide a point cloud for geometry initialization. **Bottom** (decoupled imaging): The illumination and detection paths are separated by routing the laser through an independent collimator. This mode captures multi-bounce, time-resolved light transport under varying illumination conditions, forming the basis of plenoptic transient measurements used for inverse transient rendering and applications.

multiple indirect reflections—tasks that pose significant difficulties for existing methods.

2 RELATED WORK

The plenoptic function parameterizes all visual information in a scene as a seven-dimensional radiance field over spatial position, viewing direction, wavelength, and time, and it underpins models of image formation and light transport [Adelson and Bergen 1991]. The classical rendering equation characterizes steady-state light transport and can be interpreted as computing time-integrated slices of this function [Kajiya 1986]. More recently, the transient rendering equation generalizes the steady-state formulation to time-resolved transport by explicitly accounting for the finite speed of light [Smith et al. 2008]. Since the late 1990s, research on light transport has gradually evolved from early theoretical modeling towards practical acquisition and computational reconstruction, leading to a broad range of imaging and computational techniques.

2.1 Imaging Systems

Imaging systems are typically designed with specific sensor configurations and physical architectures to sample particular dimensions of the plenoptic function. Rather than improving spatial resolution alone, multi-view systems, using synchronized camera arrays or moving cameras, capture high-resolution information across spatial positions and viewing directions under approximately fixed illumination [Schönberger and Frahm 2016; Wilburn et al. 2005]. To densely sample the angular component of radiance, light-field cameras employ microlens arrays or camera arrays to capture light fields [Gortler et al. 1996; Levoy and Hanrahan 1996; Ng et al. 2005; Wu et al. 2017]. These two paradigms converge when both position and direction are sampled densely. From the reflectance perspective, multi-light (photometric) systems fix the camera and vary illumination conditions to sample the appearance domain over spectra and lighting directions [Hertzmann and Seitz 2005; Sato et al. 2003; Woodham 1980]. More recently, light-stage systems integrate dense, calibrated multi-view camera arrays with programmable, multi-source illumination inside controlled enclosures [Debevec 2012; Debevec et al. 2000; Guo et al. 2019]. By design, these systems treat light transport as steady-state, integrating radiance over the exposure.

Time-resolved imaging systems sample the temporal dimension of the plenoptic function at picosecond scales and beyond. A large body of work has visualized light in flight as it propagates through a scene [Baek and Heide 2021; Gariepy et al. 2015; Malik et al. 2023b; O'Toole et al. 2017; Velten et al. 2013]. Temporal resolutions ranging from picoseconds down to femtoseconds have been achieved using techniques such as holography [Abramson 1978], interferometry [Gkioulekas et al. 2015; Kotwal et al. 2023], and streak cameras [Feng et al. 2022; Liang et al. 2014; Velten et al. 2013], often in combination with ultrafast pulsed lasers. Recently, single-photon avalanche diodes (SPADs) have become prevalent due to their favorable balance of cost and high temporal resolution, particularly for non-line-of-sight imaging, including looking around corners [Buttafava et al. 2015; Lindell et al. 2019; Liu et al. 2019; O'Toole et al. 2018; Shen et al. 2021; Xin et al. 2019; Ye et al. 2021] and imaging through scattering media [Lindell and Wetzstein 2020; Satat et al. 2018; Wang et al. 2023]. Whether using a scanning single-pixel SPAD or a SPAD array, these systems typically observe the scene from a single viewpoint, which limits spatial and angular resolution.

Malik et al. [2024] introduce the first SPAD-based platform for multi-view transient imaging, independently actuating the laser and detector while co-moving the illumination with the scene. Their follow-up work [Malik et al. 2025] couples the light source to the SPAD and instead rotates the scene independently. Although these setups aim to span the full dimensionality of the plenoptic function, the source–sensor coupling and reliance on mechanical motion limit truly independent control of illumination, viewpoint, and scene, which constrains fine-grained decomposition. In contrast, GenPIE provides separate actuation of the SPAD, the laser, and the scene, effectively serving as a light-transport stage. This design enables comprehensive sampling of the plenoptic function across

spatial, temporal, and angular dimensions and yields time-resolved measurements decoupled from viewing and illumination directions.

2.2 Processing Light Transport Measurements

Once captured, light-transport data can be processed via inverse rendering to recover a scene's physical attributes, such as geometry, material reflectance, and lighting. Since this inverse problem is highly ill-posed, classical methods rely on simplified physics (e.g., Lambertian or single-bounce assumptions) and/or strong priors [Kang et al. 2006]. In multi-view setups, light-field techniques synthesize novel views by parameterizing radiance in a 4D ray space [Levoy and Hanrahan 1996; McMillan and Gortler 1999]; the lumigraph further incorporates depth (or proxy geometry) to reduce rendering ambiguity [Furukawa and Ponce 2010; Gortler et al. 1996; Schönberger and Frahm 2016]. In parallel, photometric methods often presume known (or geometrically simple) scene structure to estimate reflectance parameters [Barron and Malik 2014; Zhang et al. 1999]. These paradigms typically keep either the spatial or the angular domain fixed while estimating scene properties along the other dimension. With high-dimensional measurements from light-stage systems, geometry, illumination, and spatially varying BRDFs can be jointly estimated [Dana et al. 1999; Guo et al. 2019]. More recently, differentiable and neural inverse-rendering methods—most notably neural radiance fields (NeRF)—use learned priors to reduce reconstruction ambiguity and improve the estimation of scene attributes [Ge et al. 2025; Mildhall et al. 2021; Srinivasan et al. 2021; Zhang et al. 2021]. A growing trend is integration of foundation-model priors, e.g., DINOv3 [Siméoni et al. 2025] and CLAY [Zhang et al. 2024], aiding the recovery of high-resolution 3D geometry and appearance from sparse views [Xiang et al. 2025, 2024]. Concurrently, feed-forward 3D foundation models, for instance, VGGT [Wang et al. 2025] and MASt3R [Leroy et al. 2024], learn to directly infer consistent 3D scenes. These methods demonstrate remarkable robustness, and enables high-fidelity reconstruction even from sparse multi-view observations.

From time-resolved measurements, the core inverse problem is to recover the distribution of photon path lengths. A primary application lies in 3D scene reconstruction [Gkioulekas et al. 2015; Kirmani et al. 2014; Shin et al. 2016; Zhou et al. 2015]. Since the transient profile carries detailed signatures of both direct and indirect light transport, numerous studies explicitly separate these components to improve shape and material recovery [Klinghoffer et al. 2024; Lindell et al. 2018; Malik et al. 2025; O'Toole et al. 2014; Wu et al. 2014]. A classic example of inverse transient rendering is non-line-of-sight reconstruction. Relying solely on indirect reflections (often captured via a relay wall), transient information significantly improves the conditioning of the inverse problem compared to intensity-only measurements. This enables the recovery of geometry and albedo around a corner [Ahn et al. 2019; Lindell et al. 2019; Liu et al. 2019, 2023; O'Toole et al. 2018; Xin et al. 2019], or even around two corners [Royo et al. 2023b]. Recent NLOS approaches increasingly combine differentiable transient light transport models with learned priors or neural scene representations to better handle noise, sparsity, and complex materials [Chen et al. 2020; Fujimura et al. 2023; Li et al. 2023; Mu et al. 2025; Shen et al. 2021]. Beyond looking around corners, inverse transient rendering also underpins imaging in scattering media (e.g., fog, underwater, and tissue-like media), capable of depth estimation in participating media and reconstruction under strong backscatter [Du et al. 2022; Lindell and Wetzstein 2020; Wang et al. 2023].

Many efforts have been made to reconstruct light transport from multi-view and time-resolved measurements [Feng et al. 2022; Malik et al. 2025, 2024, 2023a; Mu et al. 2024]. These inverse transient rendering methods, however, require substantial time for data acquisition and processing due to the dense sampling across spatial, temporal, and angular dimensions. To address this, GenPIE integrates a foundation model into the inverse transient rendering framework. This strategy provides a strong geometric prior even from sparse measurements and helps resolve ambiguities. More importantly, this physically grounded prior allows our method to go beyond the coarse separation of direct and indirect components from the time-resolved measurements commonly adopted in previous work. Our technique can disentangle direct and multi-bounce indirect components and recover physically grounded light transport.

3 TRANSIENT RENDERING EQUATION

3.1 From Steady-State to Transient Rendering

The classical rendering equation models outgoing radiance $L_o(\mathbf{x}, \omega_o)$ at a surface point $\mathbf{x}$ assuming steady-state equilibrium, where the radiance field remains temporally invariant [Immel et al. 1986; Kajiya 1986]:

$$L_o(\mathbf{x}, \omega_o) = L_e(\mathbf{x}, \omega_o) + \int_\Omega f_r(\mathbf{x}, \omega_i, \omega_o)\, L_i(\mathbf{x}, \omega_i)\, (\omega_i \cdot \mathbf{n})\, d\omega_i, \quad (1)$$

where f_r is the bidirectional reflectance distribution function (BRDF), L_i is incident radiance from a direction ω_i, L_e is emitted radiance along direction ω_o, and $\mathbf{n}$ is the surface normal at $\mathbf{x}$. The integration domain Ω represents the hemispherical space of incident directions above the surface. This formulation implicitly integrates light transport over time and therefore conflates contributions from light paths of different lengths, making it unsuitable for reasoning about the temporal structure of photon propagation.

Recent advances in transient imaging resolve photon arrival times with picosecond-level accuracy, necessitating an explicit temporal formulation of light transport.

Let $L_o(\mathbf{x}, \omega_o, t)$ denote the outgoing radiance at location $\mathbf{x}$, direction ω_o, and time t. The transient rendering equation extends the steady-state formulation Eq. (1) by explicitly accounting for the finite speed of light and propagation delays along light paths [Smith et al. 2008]:

$$L_o(\mathbf{x}, \omega_o, t) = L_e(\mathbf{x}, \omega_o, t) + \int_\Omega \int_0^t f_r(\mathbf{x}, \omega_i, \omega_o) L_i(\mathbf{x}, \omega_i, t-\tau) \cdot (\omega_i \cdot \mathbf{n}) \delta(\tau - d/c)\, d\tau\, d\omega_i. \quad (2)$$

Here, the optical path length d is defined as the Euclidean distance between the surface point $\mathbf{x}$ and the previous interaction point $\mathbf{x}'$ along the incident direction $-\omega_i$ and c is the speed of light. The Dirac delta function restricts the integration to light paths sharing the same total optical length. This enables the separation of direct

illumination (earliest arrival) from global light transport. However, due to the complexity of scene geometry and reflectance, the angular integration over Ω persists. At any specific time t, the outgoing radiance represents an aggregate of distinct multi-bounce paths that coincidentally share the same propagation delay ($\tau = d/c$). Thus, Eq. (2) models the signal not as fully disentangled paths, but as time-resolved integrals. Nevertheless, it still imposes stricter constraints than steady-state rendering on the admissible light transport paths by conditioning them on the optical path lengths.

3.2 Inverse Transient Rendering

In Eq. (2), the time-resolved radiance $L_o(\mathbf{x}, \omega_o, t)$ describes directional photon flux at scene level. However, practical transient imaging sensors do not measure that time-resolved radiance directly. Instead, they record irradiance integrated over the sensor aperture, angular response, and finite temporal resolution. Let $I(\mathbf{p}, t)$ denote the ideal transient irradiance at pixel $\mathbf{p}$ and time t, obtained by integrating contributions from all visible surface points $\mathcal{S}$:

$$I(\mathbf{p}, t) = \iint_{\mathcal{S}} L_o(\mathbf{x}, \omega_{\mathbf{x}\to\mathbf{p}}, t)\, G(\mathbf{x}, \mathbf{p})\, V(\mathbf{x}, \mathbf{p})\, d\mathbf{x}, \tag{3}$$

where G represents geometric attenuation (including inverse-square falloff and cosine terms), and V is a binary visibility function between $\mathbf{x}$ and $\mathbf{p}$. In practice, measurements are discretized into transient histograms $\mathcal{T}_{\text{meas}}(\mathbf{p}, t_k)$ due to the system impulse response $\Psi(\tau)$ and finite temporal bin width Δt. The measured transient is given by Eq. (4):

$$\mathcal{T}_{\text{meas}}(\mathbf{p}, t_k) = \int_{t_k-\Delta t/2}^{t_k+\Delta t/2} \left[I(\mathbf{p}, \tau) * \Psi(\tau) \right] d\tau. \tag{4}$$

We formulate the inverse transient rendering problem as approximating a (pseudo-)inverse operator $\mathcal{F}_{\text{inv}} \approx \mathcal{F}_{\text{fwd}}^{-1}$ that seeks to infer scene parameters $\Theta = \{\mathcal{G}, \mathcal{M}, \mathcal{I}\}$ (geometry, materials, and illumination) that best explain the observed time-resolved measurements $\mathcal{T}_{\text{meas}}$. $\mathcal{F}_{\text{fwd}}$ denote this forward transient imaging operator, which is governed by both the Transient Rendering Equation (Eq. (2)) and the sensor measurement model (Eq. (4)), such that $\mathcal{T}_{\text{meas}} = \mathcal{F}_{\text{fwd}}(\Theta)$.

Existing approaches of inverse transient rendering primarily differ in how they parameterize or approximate the inverse operator $\mathcal{F}_{\text{inv}}$. Early systems relied on strong assumptions to mitigate the inherent ill-posedness of the inverse problem. For example, Femto-Photography [Velten et al. 2013] assumes known scene geometry and focuses on visualizing transient light propagation, thus collapsing $\mathcal{F}_{\text{inv}}$ to a restricted setting and bypassing the recovery of full scene parameters. Flying with Photons [Malik et al. 2024] realizes the inverse operator implicitly using a neural field that encodes scene geometry and is optimized to reproduce measured transients. It enables novel view rendering of propagating light but without explicitly recovering a physically interpretable decomposition into geometry, materials, and illumination. InvProp [Malik et al. 2025] takes a step towards physically grounded reconstruction from time-resolved measurements. It introduces a hybrid inverse rendering framework that models geometry via volume rendering and uses physically-based ray sampling and accounts for indirect light through a time-resolved radiance cache.

Our work, in contrast, addresses this challenging inverse problem by both enhancing measurement diversity $\mathcal{T}_{\text{meas}}$ using a dual-mode imaging system and regularizing scene parameter estimation Θ through a physically-based differentiable path tracing. Our system decouples the illumination and sensing paths, enabling the capture of both accurate sparse geometry via confocal scanning and plenoptic transients that span the spatial, temporal, angular, and illumination domains using decoupled imaging. This rich measurement diversity provides the necessary observational anchorage to resolve geometric and material ambiguities. To overcome the challenge of incomplete data, we leverage 3D foundation models to transform sparse point clouds into complete geometric priors. These parameters are then refined via a differentiable path tracer by iterating the transient rendering equation, ensuring that the final recovered attributes are consistent with the physical laws of time-resolved light propagation.

3.3 NLOS Imaging as Inverse Transient Rendering

Non-Line-of-Sight (NLOS) imaging aims to reconstruct the geometry and scattering properties of hidden scenes through indirect light transport. Within our theoretical framework, NLOS reconstruction can be specially formulated as a constrained inverse transient rendering problem.

While the general transient rendering equation (Eq. (2)) recursively models arbitrary multi-bounce light transport, NLOS imaging imposes structural constraints through carefully designed acquisition setups. In conventional NLOS imaging systems, a pulsed laser illuminates a known relay surface (e.g., a wall), which scatters light into an occluded space Ω_h. After interacting with hidden objects, photons return to the relay surface and are detected by a time-resolved sensor. The known geometry of the relay surface and calibrated propagation paths (laser-to-wall and wall-to-detector) allow explicit modeling to focus solely on hidden scene interactions.

Under this formulation, the transient response measured at detector pixel $\mathbf{p}$ and time t becomes:

$$\begin{aligned}\mathcal{T}_{\text{nlos}}(\mathbf{p}, t) = \int_{\Omega_h} & f_r(\mathbf{x}_h, \omega_{\mathbf{x}_l\to\mathbf{x}_h}, \omega_{\mathbf{x}_h\to\mathbf{x}_d}) \cdot G(\mathbf{x}_l, \mathbf{x}_h, \mathbf{x}_d) \cdot \\ & L_e(\mathbf{x}_l, \omega_{\mathbf{x}_l\to\mathbf{x}_h}) \cdot \delta\left(\frac{t}{2} - \frac{\|\mathbf{x}_l - \mathbf{x}_h\| + \|\mathbf{x}_h - \mathbf{x}_d\|}{c}\right) d\mathbf{x}_h,\end{aligned} \tag{5}$$

where Ω_h is the domain of the hidden scene, $\mathbf{x}_l$ and $\mathbf{x}_d$ denote the laser and detector positions on the relay wall, and $\mathbf{x}_h$ is a surface point in the hidden scene. The function f_r represents the hidden surface's bidirectional reflectance distribution function (BRDF), and G includes the visibility between the relay wall and hidden surfaces, the cosine foreshortening at each interaction point, and the $1/r^4$ propagation attenuation for both forward and backward paths. The Dirac delta function ensures that only photons with travel time matching the measured delay contribute to the response.

Although structurally simpler than the general recursive transient rendering equation, Eq. (5) is actually its direct specialization for three-bounce light transport. By assuming a known relay geometry, we can reduce the complex light transport to a single integral over

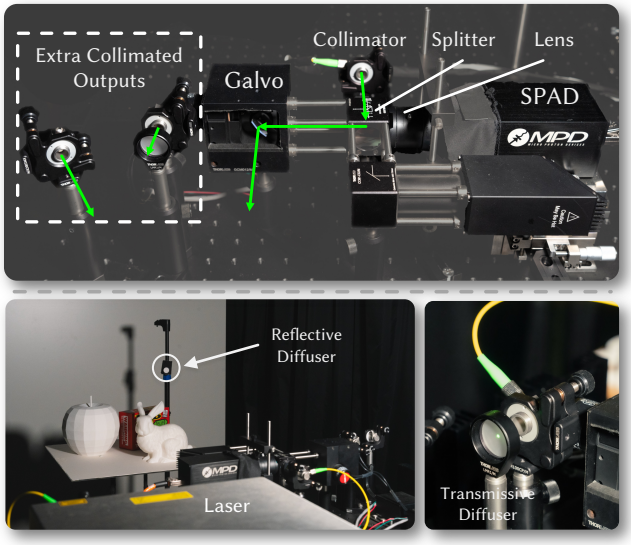

Fig. 3. Experimental hardware setup. The objects are placed on a rotation stage for changing perspectives. A pulsed laser beam is routed via an optical fiber to flexibly illuminate the scene. The beam can be steered to a reflective diffuser inside the scene (bottom left), or through a transmissive diffuser onto the scene to create a point light source (bottom right).

the hidden scene. While this simplification renders the inverse problem computationally tractable, it still has the inherent ill-posedness of transient rendering, particularly because of limited observability on the relay surface and sparse data and low signal-to-noise ratio.

4 CAPTURING PLENOPTIC TRANSIENTS

We introduce a flexible acquisition system designed to capture plenoptic light transport. Unlike prior setups, our system employs a physically reconfigurable scanning architecture. It integrates a pulsed laser, a single-photon detector, and a pair of scanning galvanometer mirrors. By synchronizing the scanning optics with a motorized rotation stage, we can arbitrarily sample the spatial, angular, and temporal dimensions of light transport.

4.1 Dual-Mode Imaging System

Our system allows independent control over lighting and viewing configurations. To illuminate the scene from arbitrary directions, we first use a collimator to convert the output free-space laser beam into a collimated beam. We then use an optical fiber to relay the beam into different components on the optical table, allowing us to flexibly change the lighting condition and control the incident direction without disturbing the overall system alignment. We place the scene on a rotation stage so that we can capture light transport from different viewpoints around the scene.

Confocal imaging. As shown in Fig. 2 (top), the optical fiber is connected to a collimator oriented toward a beam splitter. The collimated laser beam is then split and directed to a pair of galvo mirrors, which raster-scan the scene. The back-scattered light follows the same optical path in reverse, but passing through the beam splitter and being detected by a SPAD. Under this configuration, the system operates in a confocal imaging mode and functions as a single-photon lidar system where our system collects direct time-of-flight measurements from multiple viewpoints. We can then reconstruct sparse point clouds with accurate metric scale. These point clouds are then used to initialize a scene geometry given by a generative foundation model.

Decoupled imaging. Using decoupled imaging, the lighting and detection paths are no longer constrained to share the same optical axis. As shown in Fig. 2 (bottom), we decouple the lighting path by removing the fiber from the collimator facing the beam splitter and instead placing additional collimators either on the optical table or directly inside the scene. Under this configuration, the galvo mirrors only steer the detection path (SPAD's field of view). In practice, we place a collimator on the optical table and attach a diffuser in front of it, to approximate a point light source emitted from the optical table. To place the light source inside the scene, we put reflective diffusers inside the scene and directly illuminate the center of each diffuser using the collimator mounted on the optical table, thereby creating a secondary light source that illuminates the scene as a point light source.

Compared to confocal imaging, the decoupled imaging provides three distinct advantages: First, it supports time-resolved measurement of light transport in varying lighting configurations and viewpoints. Second, it captures rich global light transport phenomena such as inter-reflections and subsurface scattering that are typically suppressed in confocal setups. Finally, it provides the necessary angular diversity to constrain the Transient Rendering Equation (Eq. (2)), ensuring that the inverse problem is well-posed for estimating material and geometry properties.

4.2 Experimental Setup

System implementation. We implemented a physical prototype of the proposed architecture. The active illumination is provided by a pulsed laser *Katana 05-HP* (532 nm) operating at a 1 MHz repetition rate. For detection, we use a single-pixel Single-Photon Avalanche Diode (SPAD) from MPD, connected to a Time-Correlated Single Photon Counting module (TCSPC) *PicoHarp 300*. This imaging system achieves a 4 ps temporal resolution, with a ~72 ps time jitter. In all experiments, we apply temporal binning with a factor of 2, resulting in 16 ps temporal resolution. To enable multi-view acquisition, the target scene is placed on a high-precision motorized rotation stage (Fig. 3).

Calibration of external lighting positions. To calibrate the positions of the decoupled light sources (e.g., a reflective diffuser placed inside the scene), we utilize the confocal configuration to establish a precise mapping between galvo control voltages and physical ray directions. We manually steer the galvo mirrors via control voltages (V_x, V_y) until the laser spot is centered on the target reflector. Due to the approximately affine mapping between the galvo control voltages and the beam deflection angles, we model the angular direction of the emitted ray as

$$[\theta, \phi]^\top = \mathbf{A} \begin{bmatrix} V_x \\ V_y \end{bmatrix} + \mathbf{b}, \tag{6}$$

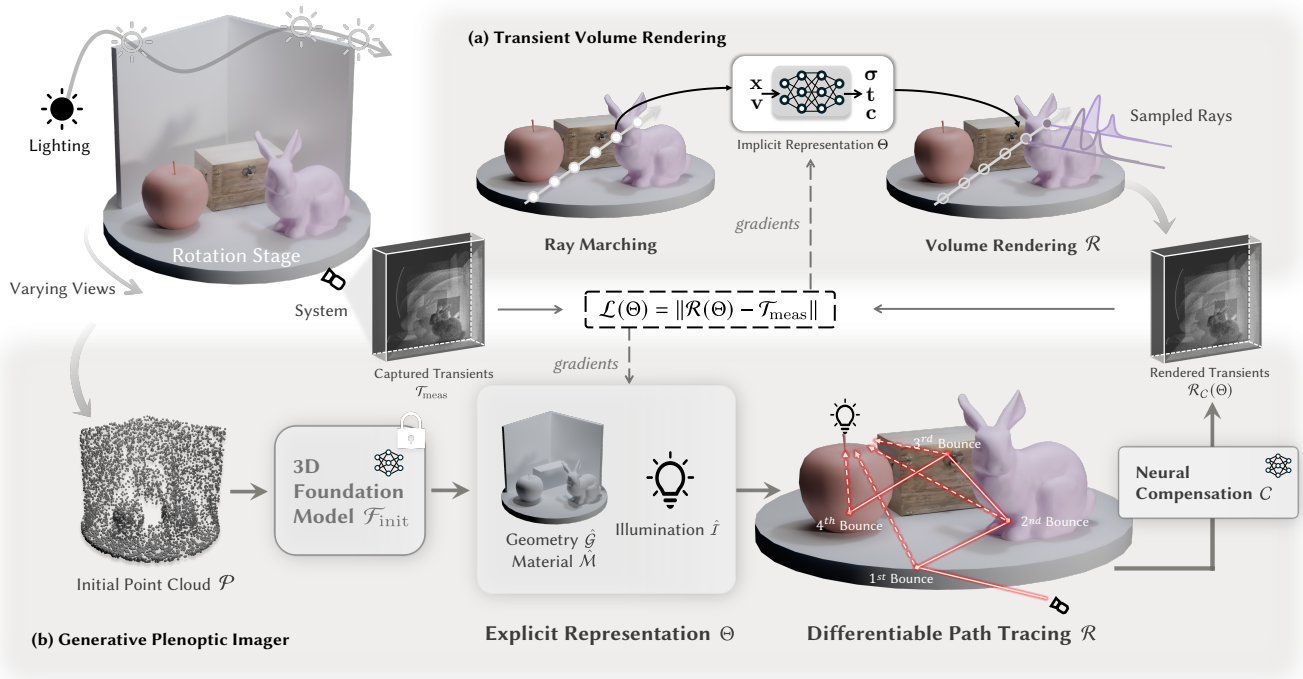

Fig. 4. Overview of our framework. **(a)** Transient volume rendering uses a neural implicit representation as scene parameters. It gives the intensity and transient response at any position inside the scene. **(b)** Because of the special design of dual-mode imaging system, our work can capture direct time-of-flight (dToF) as well as transient light propagation using a single system without extra calibration. We then use a 3D foundation model to convert the point cloud given by dToF signals into an initial geometry, which is further optimized explicitly, together with material appearance and illumination by transient rendering equation-based differentiable path tracing.

where (θ, ϕ) denote the azimuth and elevation angles in the camera coordinate system. The parameters $\mathbf{A}, \mathbf{b}$ are pre-determined by fitting a grid of known scan points. By applying Eq. (6) to the target voltages, we can recover the lighting direction in the camera frame.

Finally, we manually measure the physical distance between the collimator exit on the optical table and the center of the diffuse reflector. This distance determines the optical path length from the laser source to the reflector, which is used to calculate the position of the reflective diffuser in the world space. Other than that, we can also calculate corresponding temporal offset and shift the raw transient measurements by the appropriate number of time bins to make $t = 0$ when the beam hits the reflective diffuser.

Acquisition procedure. Our data acquisition follows a coarse-to-fine strategy. We arrange a textured scene including a checkerboard of known dimensions for scale reference.

(1) **Pose estimation**: We first capture intensity images (using accumulated photon counts) at 15° intervals. The camera intrinsic and extrinsic parameters are then jointly estimated using VGGT [Wang et al. 2025]. We rescale the extrinsic translation vectors to match the physical size of the checkerboard.

(2) **Geometric initialization:** We perform the confocal imaging at 45° intervals. These measurements are used to reconstruct an initial sparse point cloud.

(3) **Plenoptic transients acquisition:** Finally, we acquire light transport data using the decoupled imaging. We define a set of viewing angles $\mathcal{V}$ spanning a 90° arc and a set of diverse illumination positions $\mathcal{I}$. We iterate through all pairwise combinations $(\mathbf{v}, \mathbf{i}) \in \mathcal{V} \times \mathcal{I}$. For each pair, we record the transient histogram, resulting in plenoptic transient measurements.

5 GENERATIVE INVERSE TRANSIENT RENDERING

Given the plenoptic transient measurements captured by our imaging system, our goal is to solve the inverse transient rendering problem by recovering physically consistent scene parameters, including geometry, material reflectance, and illumination. To achieve this, we propose a hybrid reconstruction framework that combines data-driven geometric priors from foundation models with physically-based constraints from transient light transport. This formulation addresses two key challenges in transient scene analysis: (1) incomplete and sparse observations caused by occlusions and limited angular sampling, and (2) the need to enforce physical consistency with time-resolved measurements governed by the Transient Rendering Equation.

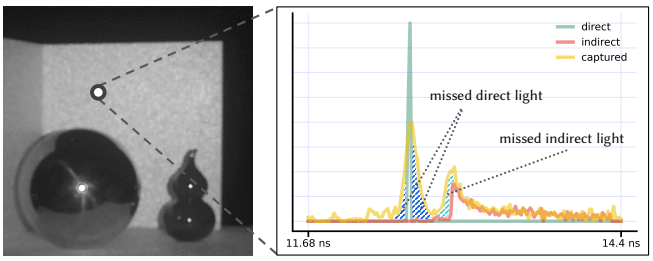

Fig. 5. The effect of system response in real transient acquisition. Ideal transient rendering (blue & green) assumes a Dirac-like impulse response, whereas real-world SPAD measurements (yellow) suffer from complex temporal broadening and jitter. This misalignment leads to errors in applications such as disentangling multi-bounce light transport.

5.1 Initialization with Foundation Model

First, we acquire a sparse point cloud $\mathcal{P}$ via confocal imaging and multi-view scans. While direct time-of-flight offers superior signal-to-noise ratios, our framework can also operate on decoupled measurements, as the first-photon arrival time remains indicative of surface geometry even under complex light transport [Malik et al. 2024].

However, $\mathcal{P}$ is typically incomplete due to self-occlusions and view limits. We thus employ a foundation model as a data-driven geometric prior that maps the sparse point cloud $\mathcal{P}$ to an initial surface mesh

$$\hat{\mathcal{G}}_{\text{init}} = \mathcal{F}_{\text{init}}(\mathcal{P}), \tag{7}$$

where $\hat{\mathcal{G}}_{\text{init}}$ denotes the reconstructed scene geometry. Importantly, the foundation model is used solely to initialize geometry. The predicted mesh $\hat{\mathcal{G}}_{\text{init}}$ serves as a structural prior that is subsequently refined through physically-based inverse transient rendering.

5.2 Optimization via Differentiable Path Tracing

The geometry generated by the foundation model provides a structural prior but lacks material properties and fine-scale details. To enforce physical consistency, we refine the scene parameters by differentiable path tracing [Yi et al. 2021], which explicitly models transient light propagation according to the transient rendering equation Eq. (2). In this stage, the optimization is performed using multi-view and multi-illumination transient videos.

Let $\Theta = \{\hat{\mathcal{G}}, \hat{\mathcal{M}}, \hat{\mathcal{I}}\}$ represent the complete set of estimated scene parameters (geometry $\hat{\mathcal{G}}$, material $\hat{\mathcal{M}}$, illumination $\hat{\mathcal{I}}$, where hats distinguish the estimates from the underlying true parameters $\mathcal{G}$, $\mathcal{M}$, $\mathcal{I}$ defined in Sec. 3.2). Unlike implicit neural representations (e.g., NeRF/SDF fields) that obscure geometric surfaces, we represent the scene geometry $\hat{\mathcal{G}}$ using explicit triangular meshes. The inverse problem is formulated as minimizing the objective function:

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_{\mathbf{v},\mathbf{l}} \sum_{\mathbf{p},t_k} \rho[\mathcal{R}(\mathbf{p}, t_k; \Theta; \mathbf{v}, \mathbf{l}) - \mathcal{T}_{\text{meas}}(\mathbf{p}, t_k; \mathbf{v}, \mathbf{l})], \tag{8}$$

where $\rho(x) = \lambda_1|x| + \lambda_2 x^2$, and $\mathbf{v}$ and $\mathbf{l}$ denote the viewing and illumination configurations under decoupled imaging, respectively. $\mathbf{p}$ indexes the sensor pixels, and t_k denotes the center of the k-th temporal bin. $\mathcal{R}(\Theta; \mathbf{v}, \mathbf{l})$ denotes the forward rendering operator

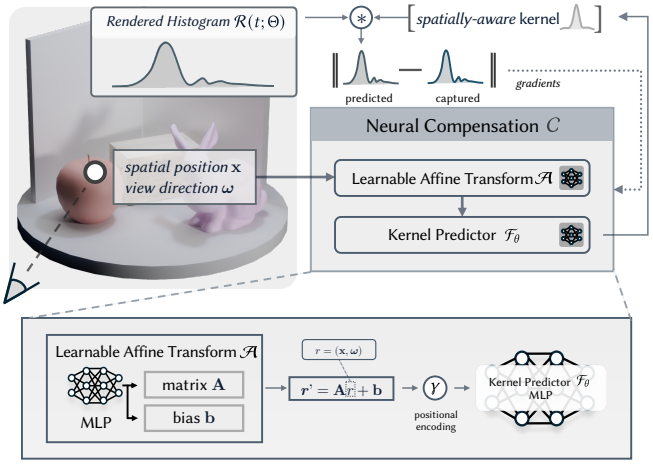

Fig. 6. Neural Compensation Pipeline. We account for system non-idealities by learning a spatially-varying kernel. For a sampled ray with spatial intersecting position $\mathbf{x}$ and view direction $\boldsymbol{\omega}$, we first predict an affine transform that maps the raw coordinates to a canonical space. The transformed coordinates are then processed via positional encoding and a kernel predictor MLP to produce a spatially-aware kernel, which is then convolved with the initially rendered transient histograms. Overall, the module acts as a linear operator on rendered transients.

that simulates transient light transport given scene parameters Θ under a specific view and illumination configuration. $\mathcal{T}_{\text{meas}}$ represents the measured transients. The outer summation enumerates all view and illumination configurations and the inner summation computes pixel-wise and time bin differences. N normalizes by the total samples.

To optimize Θ, we derive gradients using the path-integral formulation for transient transport. The gradient of the loss is estimated via Monte Carlo sampling:

$$\nabla_\Theta \mathcal{L} \approx \frac{1}{M} \sum_{i=1}^{M} \left[\delta w^*(\bar{\mathbf{x}}_i) \cdot \nabla_\Theta \left(\frac{f(\bar{\mathbf{x}}_i, \Theta)}{p(\bar{\mathbf{x}}_i)} \right) \right], \tag{9}$$

where $f(\bar{\mathbf{x}}_i, \Theta)$ denotes the contribution function of a sampled transient light path $\bar{\mathbf{x}}_i$ under Θ, and $p(\bar{\mathbf{x}}_i)$ is the corresponding path sampling density. The path contribution $f(\bar{\mathbf{x}}_i, \Theta)$ is accumulated into the temporal bin corresponding to its total time-of-flight $\tau(\bar{\mathbf{x}})$. The adjoint weight $\delta w^*(\bar{\mathbf{x}})$ is strictly time-dependent:

$$\delta w^*(\bar{\mathbf{x}}) = \frac{\partial \mathcal{L}}{\partial \mathcal{R}(\mathbf{p}, \tau(\bar{\mathbf{x}}))}. \tag{10}$$

By utilizing differentiable path tracing, we ensure time-resolved accuracy, where each simulated path contributes to the signal with precise temporal correspondence $\tau(\bar{\mathbf{x}})$. Furthermore, it enables multi-bounce decomposition, as explicit path tracing allows us to separate and reason about direct and indirect illumination components individually, as shown in Sec. 6.

5.3 Neural Compensation

While the Transient Rendering Equation (Eq. (2)) provides a rigorous physical model of light transport, it assumes ideal conditions that

do not strictly hold in real transient measurements. Real-world capture is affected by complex system factors, including the SPAD's asymmetric temporal jitter, lens optical aberrations, and calibration errors (Sec. 3). In an ideal simulation, the system response to an impulse is modeled as a Dirac delta function; however, time jitter spreads photon arrivals across neighboring bins, as illustrated in Fig. 5. Such temporal mismatch can distort multi-bounce separation that relies on accurate timing. To resolve this, we present a neural compensation module that learns the effect of these spatiotemporal factors.

A key insight of this approach is that transient light transport follows the principle where the total transient response is the linear summation of individual bounce components in the physical world. Therefore, simply using a black-box network to regress the real transient signal would violate physical linearity. Instead, we design our network to predict a spatiotemporal kernel that acts as a linear convolution operator to preserve this fundamental physical property.

More importantly, this module is not introduced merely to improve the fit to the measured transient signal, but to make the multi-bounce decomposition physically meaningful under real system responses. It enables a more faithful estimation of how different bounce components contribute to each measured time bin. Otherwise, parts of the direct and indirect transport may be misassigned or even omitted.

As we show in Fig. 6, for a pixel $\mathbf{p}$ looking at a surface point $\mathbf{x} \in \mathbb{R}^3$ with view direction $\boldsymbol{\omega} \in S^2$, our neural compensation module $\mathcal{C}$ predicts a 1D temporal kernel $\mathcal{K}$:

$$\mathcal{K}(\mathbf{x}, \boldsymbol{\omega}) = \mathcal{F}_\theta(\gamma(\mathcal{A}(\text{concat}(\mathbf{x}, \boldsymbol{\omega})))), \tag{11}$$

where the affine transform $\mathcal{A}$ follows the ray-space embedding approach [Attal et al. 2022], mapping the ray coordinates into a canonical space to compensate for errors in camera extrinsic calibration, γ is the sin-cos positional encoding [Mildenhall et al. 2020], and $\mathcal{F}_\theta$ is an MLP. With scene parameters Θ, the final compensated transient can then be written as

$$\mathcal{T}_{\text{comp}}(\mathbf{p}, t) = \mathcal{R}(\mathbf{p}, t; \Theta) * \mathcal{K}(t; \mathbf{x}, \boldsymbol{\omega}). \tag{12}$$

Beyond modeling system responses, this module plays a crucial role in the stability of our framework. As noted before, we use an explicit mesh representation to allow for direct physical control. However, optimizing explicit mesh vertices is much more difficult than optimizing continuous implicit fields, particularly under sparse transient observations. In practice, the compensation network predicts a non-causal convolutional kernel that can shift and redistribute transient energy across neighboring bins, such that it also serves as a geometric correction that reduces discrepancies between rendered and measured transient data.

5.4 Implementation Details

We use CLAY [Zhang et al. 2024] to initialize the scene geometry from sparse and incomplete measurements. For each scene, we also temporally integrate the confocal transient measurements to obtain a steady-state intensity map, which serves as another input condition for the 3D foundation model other than the point cloud. The text prompts used for geometry generation are shown in Fig. 8. Note

Fig. 7. Comparison of geometry initialization from point cloud input with [Kazhdan et al. 2006] and [Luo et al. 2025].

Table 1. Quantitative comparison of mesh smoothness and geometric fairness. We report laplacian smoothness energy (LSE) [Meyer et al. 2003] and normal consistency (NC) [Mescheder et al. 2019] comparing with poisson surface reconstruction [Kazhdan et al. 2006] and Transientangelo [Luo et al. 2025]. All meshes are simplified to 10,000 faces before evaluation. Lower LSE or higher NC indicate smoother and more regular geometry.

Methods	LSE ($\times 10^{-5}$)↓	NC↑
Kazhdan et al. [2006]	3.2731	0.9126
Luo et al. [2025]	11.5340	0.9273
Ours	1.8664	0.9894

that the foundation model output serves strictly as an initialization; all subsequent geometric and material details are physically refined via our differentiable transient path tracing pipeline.

We extend `mitransient` [Royo et al. 2023a, 2025, 2022], a transient renderer based on Mitsuba3 [Jakob et al. 2022], for differentiable transient path tracing. We model object materials using the principled BSDF [Burley 2012, 2015], and only optimize a subset of its parameters, including base color, roughness, specular, metallic. We set the depth of path tracing to 8 for all three scenes.

For optimization, we employ a coarse-to-fine strategy to ensure robust convergence. We begin with a warm-up stage where the model is optimized against integrated (steady-state) intensity images. This step is critical for establishing a valid geometric baseline, particularly for highly metallic and specular scenes such as *Metal Sphere*. Following the warm-up, we switch to full optimization using transients captured using decoupled imaging, minimizing a weighted combination of ℓ_1 and ℓ_2 losses (Eq. (8)) between the rendered and measured time-resolved signals. We use separate Adam optimizers for geometry and materials. For scene geometry, we adopt the latent mesh parameterization [Nicolet et al. 2021] to improve stability, with the learning rate adaptively scaled based on the object's bounding box size. Regarding the estimated light source $\hat{I}$, since its initial position is pre-calibrated (Sec. 4.2), we only fine-tune its intensity

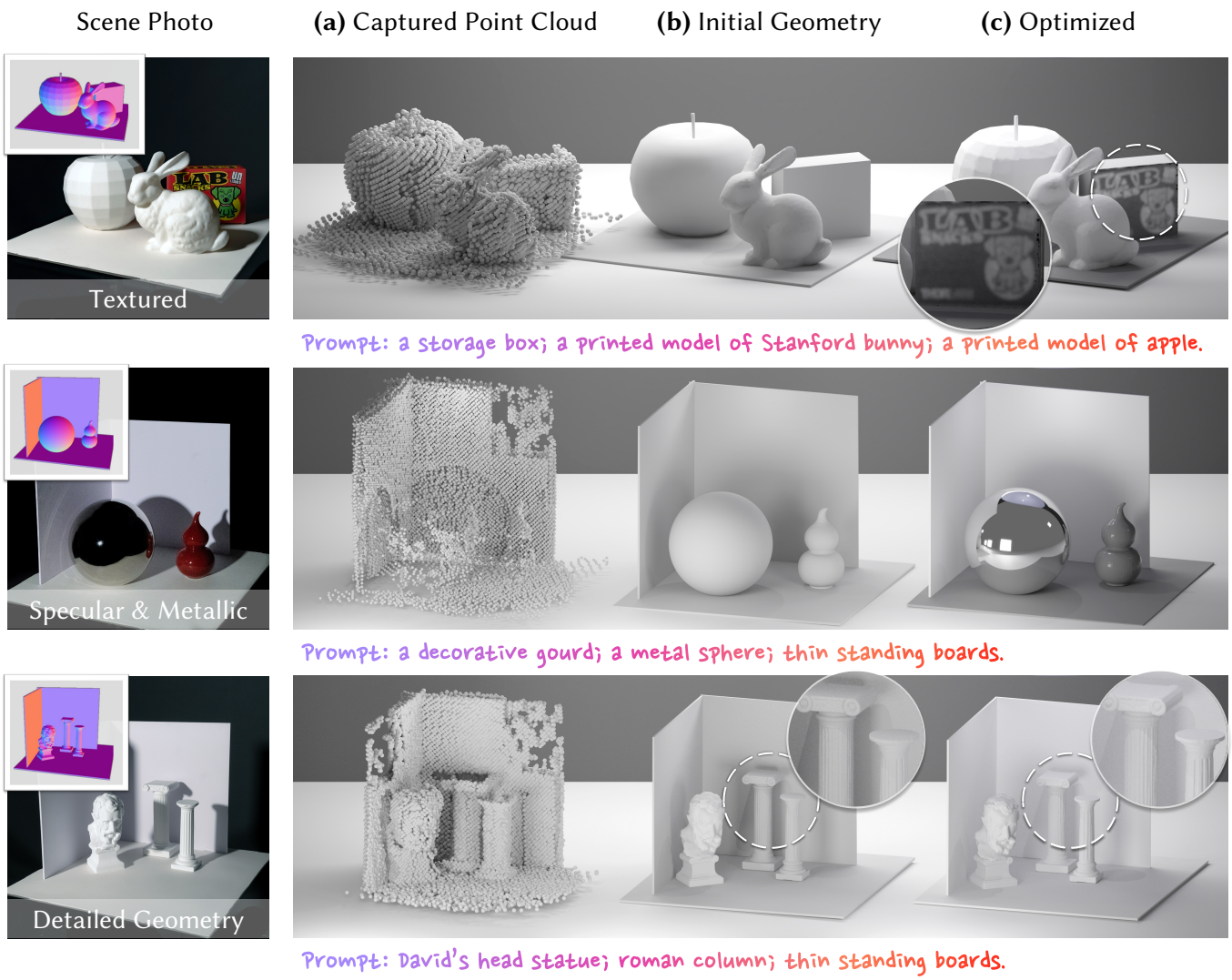

Fig. 8. We show our scene reconstruction results on different types of scene. The first column shows the conventional photos of the scenes and the normals we recover. **(a)** the captured point clouds obtained from initial transient measurements. **(b)** the initial geometry predicted from 3D foundation model used as a starting point for optimization. **(c)** the optimized results, where the scene properties, including appearances, materials, and textures are significantly refined to match the observed multi-view multi-illumination transient data. This figure is rendered using Blender by taking the optimized BSDFs of different objects.

and a small positional offset. The entire end-to-end optimization is performed on a single NVIDIA A40 (48GB) and converges within ~ 20 minutes.

We compute the spatial context $(\mathbf{x}, \boldsymbol{\omega})$ for each camera ray as input to the neural compensation module. For the *Bunny Apple Box* and *Roman Statue* scenes, we directly extract the first surface intersection $\mathbf{x}$ with the ray from the captured point clouds and derive view directions $\boldsymbol{\omega}$ from calibrated extrinsics. For the *Metal Sphere* scene, where the highly specular surface results in incomplete point cloud capture (see Fig. 8), we instead generate surface coordinates by rendering depth maps from the initialized geometry.

Table 2. Statistics of our plenoptic transient datasets. We report spatial resolution, total number of photons detected per view, and exposure time per pixel.

Scenes	Resolution ($H \times W$)	# Photons (per View)	Exp. Time (per Pixel)
Bunny Apple Box	512×512	2×10^8	30 ms
Roman Statue	512×512	2.6×10^8	30 ms
Metal Sphere	512×512	1.2×10^8	30 ms
Cola Bottle	256×512	4.7×10^7	100 ms

6 EVALUATIONS

In this section, we comprehensively evaluate GenPIE on real-world transient data. We demonstrate its capability to jointly recover high-fidelity scene geometry and material properties while effectively compensating for the physical non-idealities of the imaging system. We first assess the accuracy and topological quality of our geometry reconstruction against representative explicit and implicit baselines (Sec. 6.1). Then, we demonstrate the joint recovery of geometry and appearance, showing the refinement from initialization to final physically grounded results (Sec. 6.2). Finally, we evaluate the generalization capability of our method, especially the proposed Neural Compensation module (Sec. 6.3), via novel lidar view synthesis.

Captured dataset. To evaluate the geometric reconstruction capabilities of our approach, we first conduct lidar scan around three individual objects (Fig. 7). Beyond these single objects, we further capture several scenes for more comprehensive evaluations and to demonstrate applications across diverse materials and geometric properties: (1) textured (*Bunny Apple Box*) (2) complex reflectance (*Metal Sphere*), and (3) detailed geometry (*Roman Statue*). Each scene is within an approximate bounding volume of $70 \times 70 \times 70$ cm^3. For these scenes, we first capture confocal transient measurements every 45° as described in Sec. 4. Subsequently, we acquire decoupled transient measurements to record light transport within the scene from 7 viewpoints spaced approximately every 15° with 3 distinct illumination positions per view. Finally, we demonstrate more applications by capturing transient videos of light passing through a cola bottle. The details of these datasets are summarized in Tab. 2.

6.1 Geometry Reconstruction

We evaluate the robustness of our geometry reconstruction given point cloud inputs by comparing it against two baselines: Poisson Surface Reconstruction [Kazhdan et al. 2006], a representative approach to produce meshes from point clouds and Transientangelo [Luo et al. 2025], a state-of-the-art implicit method for SPAD-based lidar-view 3D reconstruction.

As shown in Fig. 7, we reconstruct objects with varying scales and reflectance properties. Poisson Surface Reconstruction is highly sensitive to the sparsity of the input lidar-view transient and fails to close holes in regions with low sampling density. Transientangelo produces much smoother surface than Poisson Surface; however, it still struggles with intricate geometric regions, such as the winding body of the dragon. In contrast, our pipeline provides the cleanest surface boundaries and best geometric quality across all objects. Furthermore, we give evaluations on non-referenced mesh quality metrics in Tab.1.

6.2 Scene Recovery

We further evaluate the capability of GenPIE to jointly recover scene-level geometry and surface appearance. Fig. 8 illustrates the reconstruction results across three representative scenes. As observed in column (a), the raw point clouds measured from confocal scanning are noisy and typically incomplete. This data sparsity is particularly evident in the *Metal Sphere* scene (second row), where specular reflection prevents the capture of back-scattered signals, resulting in large missing regions in the point cloud. By leveraging the 3D Generative Foundation Model conditioned on coarse text prompts, intensity maps, and point clouds, we obtain a topologically complete initial geometry shown in column (b). However, these generative priors are sometimes over-smoothed and lack specific surface properties, serving only as a geometric starting point.

Starting from this initialization, our differentiable transient path tracing pipeline effectively recovers details and material properties as shown in column (c). For the *Bunny Apple Box* scene (top row), our method successfully recovers high-frequency textures; the zoomed-in region shows that the text "LAB" on the storage box, which was absent in the untextured initialization, is clearly resolved in the final result. In the *Metal Sphere* scene (middle row), although initialized as a rough white surface, the final optimization accurately exhibits the high specularity and metallic properties of the sphere, while also fitting the specular appearance and intensity of the decorative gourd. Finally, in the *Roman Statue* scene (bottom row), the generative model initially predicts a rough and gray Roman statue and columns. Driven by the transient measurements, our optimization accurately recovers the plaster-like materials. These results validate that our approach recovers a set of physically-grounded, realistic scene parameters, which serve as a foundation for further applications.

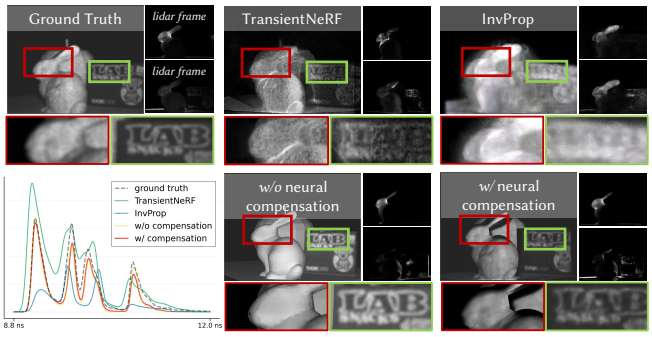

Fig. 9. Comparison of novel view synthesis. From highly sparse measurements, our framework reconstructs fine geometry and textures (e.g., "LAB" text), while TransientNeRF and InvProp suffer from volumetric blurs. The spatially-integrated transient histogram (bottom-left) and zoomed-in crops demonstrate that our approach, particularly with neural compensation, aligns the recovered light transport with the ground truth.

Table 3. Quantitative comparison on novel lidar view synthesis.

Methods	PSNR↑	SSIM↑	LPIPS↓	T-IoU↑
TransientNeRF [2023a]	19.7711	0.6165	0.5526	0.3269
InvProp [2025]	19.9379	0.3148	0.4443	0.1858
w/o Neural Compensation	20.8041	0.7758	0.2134	0.4986
w/ Neural Compensation	23.4370	0.8652	0.1769	0.6609

6.3 Novel Lidar View

We evaluate the generalization capability of our framework, especially the neural compensation module, by synthesizing novel

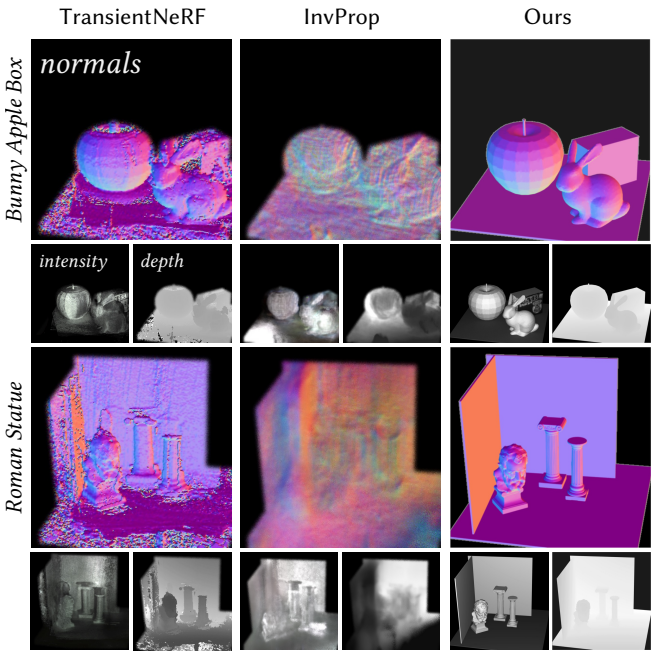

Fig. 10. Results of intensity maps, depth, and surface normals recovery from a novel view. Our method reconstructs smoother surfaces in occluded regions through an explicit mesh representation, compared to baseline methods.

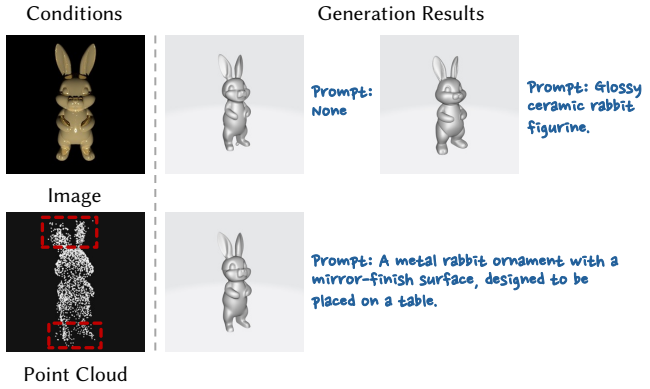

Fig. 11. We evaluate foundation-model initialization on a simulated metallic bunny scene using different prompts. The input image and sparse point cloud are shown on the left, and the generation results under no prompt, a generic glossy ceramic rabbit prompt, and a detailed metallic rabbit prompt are shown on the right. The red rectangular on the left side shows the parts of point cloud not matching the original geometry caused by inter-reflection.

transient lidar views. For this experiment, the lighting setup is consistent with that of InvProp [Malik et al. 2025]. We utilize a highly sparse setup with only 6 training views, which poses a big challenge for existing reconstruction methods.

Comparisons. We compare GenPIE against TransientNeRF [Malik et al. 2023a] and InvProp [Malik et al. 2025]. The qualitative comparisons are shown in Fig. 9. The sparse acquisition setup exposes fundamental limitations in prior works. As a physically-based inverse rendering framework, InvProp models light transport by learning a secondary ray distribution over scene geometry. While this constraint is effective under dense sampling, in our sparse setting the available physical constraints are insufficient to restrict the solution space, making it difficult for the model to consistently explain the measurements. As a result, InvProp fails to recover a complete surface, leading to fog-like volumetric artifacts. TransientNeRF, on the other hand, employs a neural field with strong capabilities to fit and interpolate high-frequency signals. Although this enables the reconstruction of detailed geometry, it also makes the model highly sensitive to small calibration errors. Under slight pose misalignments, fitting high-frequency transient signals leads to inconsistencies across views. Rather than averaging out these errors, the model overfits the view-dependent misalignment, resulting in noticeable overlapping artifacts. In contrast, our approach overcomes these limitations by combining a robust geometric initialization via the foundation model with a Neural Compensation module that accounts for discrepancies between physical simulation and sensor responses.

To assess physical fidelity, we also compare global temporal responses by spatially integrating transient measurements over the image plane ($H \times W$), as shown in Fig. 9 (bottom left). This demonstrates that our method provides a more physically grounded fit to the measured light transport compared to prior works. This fit serves as a foundation for applications such as disentangling multi-bounce light components. Accordingly, we summarize the quantitative results in Tab. 3, reporting Transient IoU (T-IoU) [Malik et al. 2024] to evaluate temporal alignment with real measurements as well as PSNR, SSIM [Wang et al. 2004], and LPIPS [Zhang et al. 2018] metrics on integrated intensity images.

Furthermore, the reconstruction of non-line-of-sight regions further highlights the robustness of our framework. As illustrated in the normal and intensity maps (Fig. 10), TransientNeRF struggles to recover complete structures in occluded areas. For example, in the *Bunny Apple Box* scene, it suffers from intensity voids on the box surface and exhibits noisy, fragmented normals in some regions. The 3D foundation models help us accurately reconstruct these challenging regions and produce a more reliable guidance of inter-reflection light transport inside scenes.

6.4 Ablation Study on Prompt Stability

We evaluate the stability of the foundation model initialization under different prompts. We conduct a simulated experiment using a metallic rabbit, where we scan a sparse point cloud and feed it to the foundation model with different text prompts.

Fig. 11 shows that generic prompts and no prompt still produce consistent initial geometry for the overall object structure. Although specular materials make the sparse point cloud more ambiguous since there are some inter-reflection between surfaces of the object, the shape are still complete.

Fig. 12. Separation of direct and indirect light transport under different illuminations, and from different viewpoints. While InvProp relies on volumetric rendering and implicit neural caches , our framework leverages explicit geometry and the Transient Rendering Equation to rigorously disentangle light paths. By enforcing physical consistency, we achieve a sharp separation, particularly between direct metallic specular reflections and indirect scattering effects on the back walls.

7 APPLICATIONS

Leveraging the recovered explicit geometry and surface properties, we extend our framework to computational tasks that require physically grounded recovery over light transport. Beyond static reconstruction, these scene parameters enable us to analyze and manipulate the captured time-resolved signals. We demonstrate this capability through four applications: (1) disentangling multi-bounce light transport components, (2) visualizing light propagation via time unwarping, and (3) time-resolved scene relighting.

7.1 Disentangling Multi-Bounce Light Transport

We disentangle multi-bounce light transport by leveraging the correspondence between our recovered scene parameters and the captured transient signals. Specifically, we first use the recovered scene parameters Θ to render transient videos $\mathcal{R}^d(\mathbf{p}, t; \Theta)$ (where d is a specific path tracing recursion depth) followed by neural compensation C to simulate transient light transport. To isolate the n-th bounce component in the synthetic domain, we render two transient videos: one up to n bounces, $\mathcal{R}^n_C(\Theta)$, and another up to $n-1$ bounces, $\mathcal{R}^{n-1}_C(\mathbf{p}, t; \Theta)$, where $\mathcal{R}_C$ denotes the rendered transient after applying the trained neural compensation module C. The isolated contribution of only the n-th bounce $\Delta\mathcal{R}_n$ is subtracted by:

$$\Delta\mathcal{R}^n(\mathbf{p}, t) = \mathcal{R}^n_C(\mathbf{p}, t; \Theta) - \mathcal{R}^{n-1}_C(\mathbf{p}, t; \Theta). \tag{13}$$

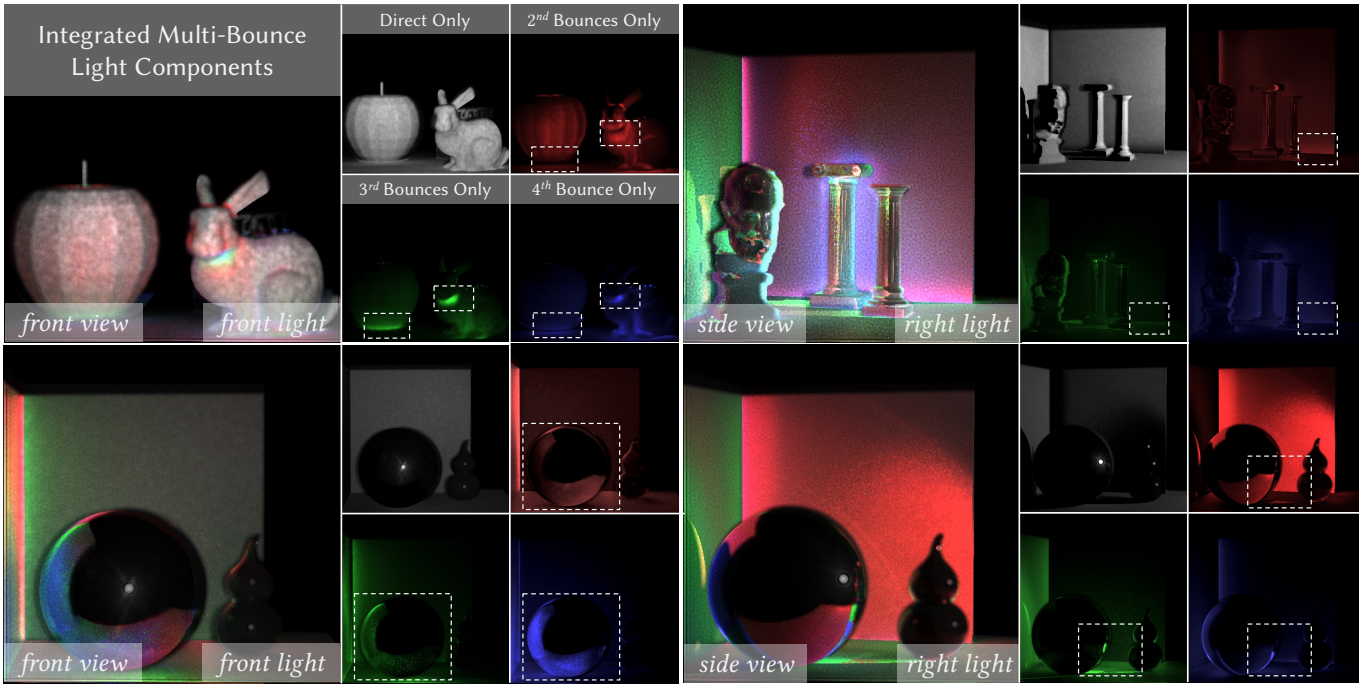

Fig. 13. GenPIE decomposes light transport components across various scenes, viewpoints, and illuminations. We visualize the direct, 2^{nd}, 3^{rd}, and 4^{th} order bounces using white, red, green, and blue, respectively, and also show their integration into a single image. Our method decomposes subtle inter-reflections inside the scenes, such as the light scattering back and forth between the apple's surface and floor.

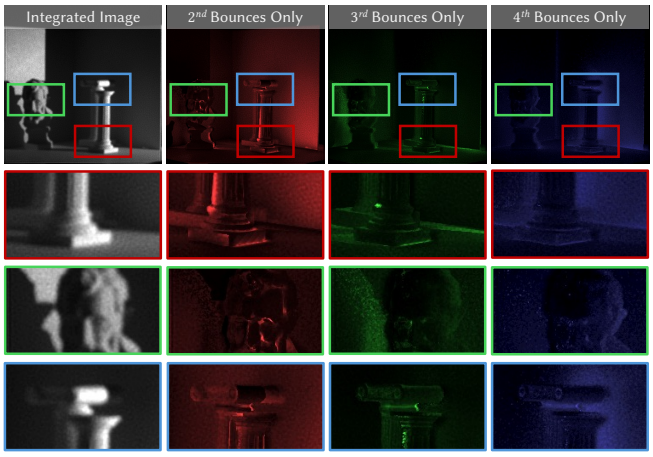

Fig. 14. More results on decomposing multi-bounce light propagation.

For each pixel $\mathbf{p}$ and time bin t, we calculate a scaling factor $\Gamma_n(p,t)$ by normalizing the isolated bounce against the total rendered radiance (up to a maximum depth D):

$$\Gamma_n(\mathbf{p},t) = \frac{\Delta\mathcal{R}_n(\mathbf{p},t)}{\mathcal{R}_C^D(\mathbf{p},t;\Theta)+\epsilon}, \tag{14}$$

where ϵ is a small constant to prevent division by zero. The final disentangled physical component $\hat{\mathcal{T}}_n$ is reconstructed by multiplying the calculated ratio with the original captured data $\mathcal{T}_{\text{meas}}$:

$$\hat{\mathcal{T}}_n(\mathbf{p},t) = \mathcal{T}_{\text{meas}}(\mathbf{p},t)\cdot\Gamma_n(\mathbf{p},t). \tag{15}$$

Comparison on direct/indirect separation. We first evaluate the separation of direct ($n=1$) and indirect ($n>1$) light components. We compare our method against a neural inverse rendering method InvProp [Malik et al. 2025], which attempts to separate direct and indirect paths based on implicit radiance caches. The results are shown from training views of InvProp. As shown in Fig. 12, InvProp struggles to combine accurate physical light transport with neural networks. For example, in the *Metal Sphere* scene, InvProp fails to cleanly isolate the specular reflection, creating foggy artifacts in the indirect component. In contrast, our physics-driven approach utilizes the explicit geometry and estimated materials to correctly classify the light paths, resulting in a clean separation where the specular reflections are confined strictly to the direct component, and the indirect light is also clearly separated.

Higher-order bounce decomposition. A key application of our framework is the ability to resolve light transport beyond only direct/indirect decomposition. Fig. 13 and Fig. 14 visualize the decomposition of transport into 2^{nd}, 3^{rd}, and 4^{th} order bounces. The 2^{nd} bounce (red) primarily captures the immediate floor-to-object inter-reflections, illuminating the base of the objects. The decomposition of higher-order components (3^{rd} and 4^{th}, green and blue) provides a visualization of the light's causal history. In the *Bunny Apple Box* scene, for example, the 2^{nd} and 3^{rd} bounces explicitly show mutual reflections between the apple's base and the floor. In the *Metal Sphere*

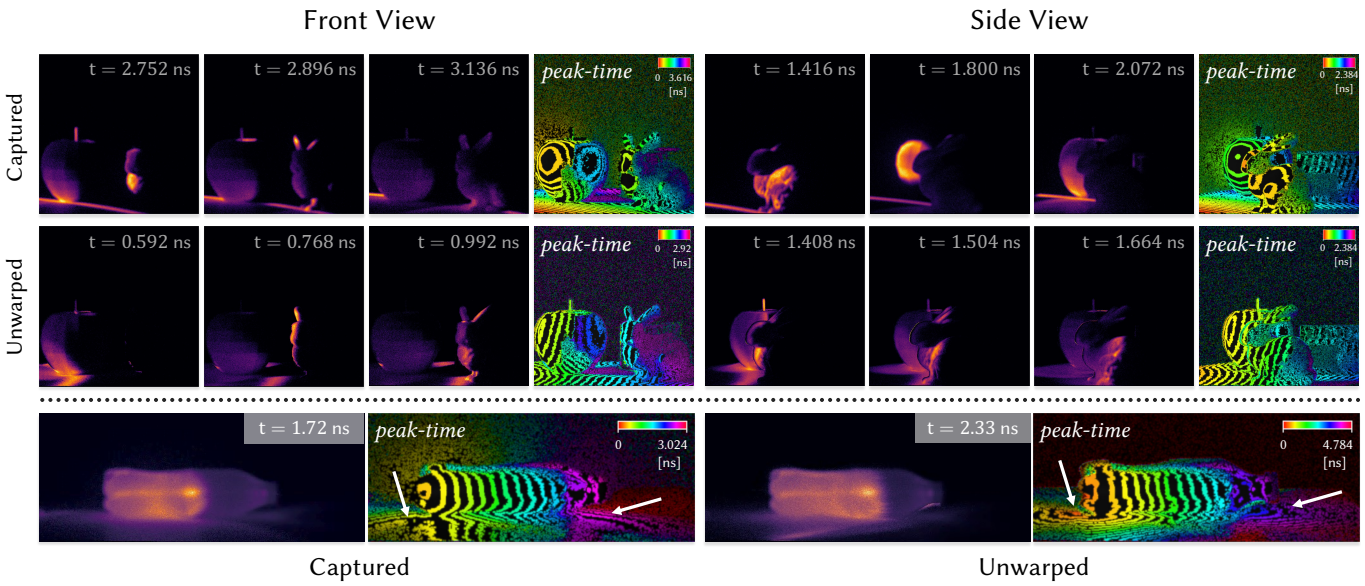

Fig. 15. Visualization of Light-in-Flight via Time Unwarping. We show several transient video frames and peak-time maps before (Captured) and after (Unwarped) applying time unwarping. The peak-time maps using the rainbow color coding scheme from Femto-photography [Velten et al. 2013] to visualize the wavefront progression. **Row 1-2** (Bunny scene, illuminated from left): After unwarping, the temporal frame shifts to the surface interaction time. **Bottom** (Cola Bottle): A similar correction is observed on the floor plane, where the camera-dependent delay is removed (indicated by white arrows).

scene (Fig. 13, bottom right), we observe that the 2^{nd} bounce primarily concentrates on the back wall, whereas the 3^{rd} bounce shifts to the left wall, also demonstrating an inter-reflection inside the scene. These components are typically blended in the raw video, and our framework successfully decodes the scattering history of light and transforms it into a structured breakdown of light propagation previously attainable only in pure simulation.

7.2 Time Unwarping

Time unwarping transforms the captured photon arrival time t into a local frame t' relative to the correct moment of surface interaction, through which we can visualize "light-in-flight" as it propagates through the scene. Formally, for a pixel $\mathbf{p}$ observing a surface point $\mathbf{x}$, the unwarped time is defined as $t' = t - \|\mathbf{x} - \mathbf{x}_c\|/c$, where $\mathbf{x}_c$ is the camera center and c is the speed of light in the corresponding medium.

Accurate unwarping relies entirely on precise depth estimation. Early pioneering work, such as Femto-photography [Velten et al. 2013], use an external digitizer arm [Faro Technologies Inc. 2012] to point-wise calibrate the scene geometry, a process that is both intrusive and slow. More recent approach like *Flying with Photons* [Malik et al. 2024] attempts to infer depth implicitly from the learned transient field; however, the unwarping quality becomes strictly bound to the accuracy of the neural reconstruction. Beyond visualization, such temporal re-alignment is critical for non-confocal NLOS imaging algorithms (e.g., Eq. (5)), which mandate that $t = 0$ matches precisely to the laser pulse impacting the relay surface. To satisfy this, current NLOS systems [Liu et al. 2020, 2019] often employ a secondary confocal SPAD setup solely for geometric calibration, escalating system costs and introducing inter-system registration errors.

Our framework solves this task simply within a single-SPAD setup by leveraging its dual-mode imaging capability (Sec. 4.1). For general (diffuse) scenes, we first perform a rapid confocal lidar scan to measure the depth map, directly establishing the temporal offsets for the scene surface. This ensures that the time-of-flight measurements in the subsequent decoupled capture are intrinsically registered to the physical geometry without external hardware. For scenes with highly metallic or specular surfaces, where confocal scanning fails to capture the complete surfaces (e.g., the *Metal Sphere*), we utilize the initialized mesh from our reconstruction pipeline (Fig. 8b) to render the depth map.

We demonstrate the robustness of this integration in Fig. 15. The *Captured* rows visualize the raw transient data, where the photon arrival times are distorted by the varying optical path lengths between the camera and the scene geometry. The peak-time visualization effectively collapses the temporal evolution of the transient video into a single frame, where the concentric isochrones reveal the radial progression of the light wavefront from the sensor's perspective. After applying our unwarping based on the calibrated geometry, the true light propagation becomes visible. In the *Bunny Apple Box* scene, the unwarped frames reveal the planar light wavefront sweeping across the objects from left to right. Similarly, the unwarping corrects the visualization of light propagating through the scattering liquid inside a cola bottle in Fig. 15 (bottom), and transforms the peak-time map into concentric isochrones centered on the bottle's caps and base, which accurately trace the diffuse reflections from the floor.

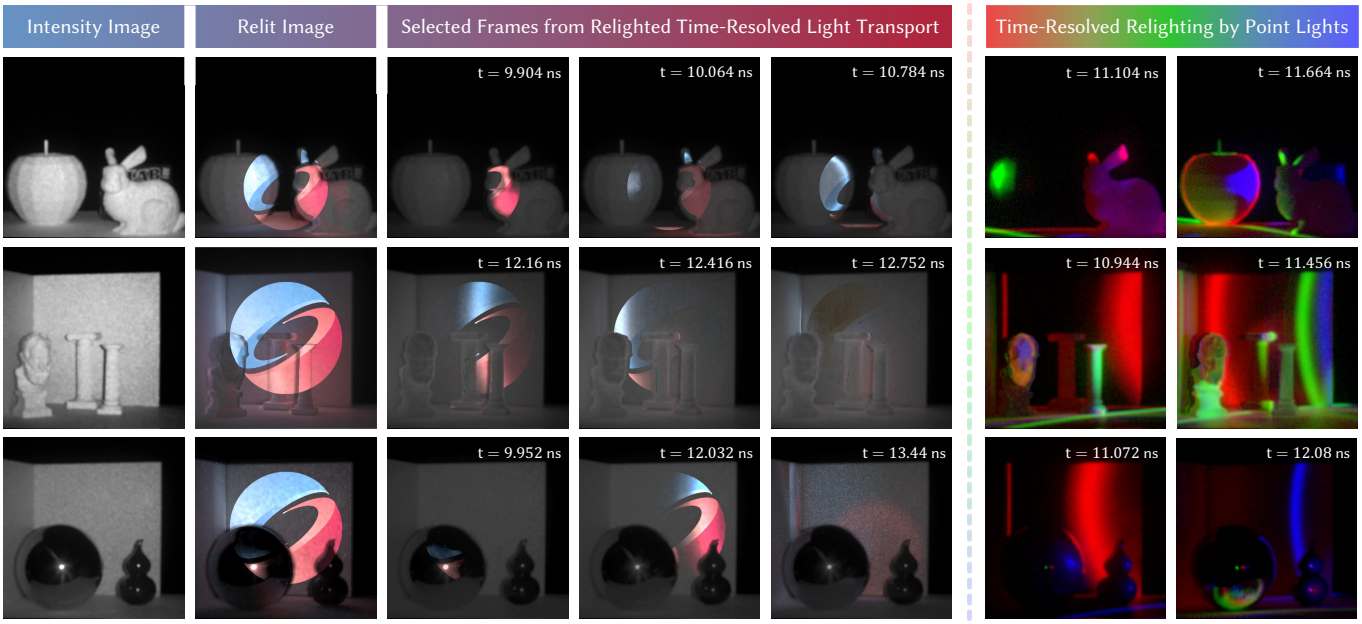

Fig. 16. Time-Resolved Relighting Results. We utilize the reconstructed geometry and materials to simulate light transport under novel illumination. **Left:** We project a virtual light source with a colored SIGGRAPH logo pattern. **Column 2** shows the integrated result, while the selected frames visualize the structured wavefront propagating through the scene, matching occlusion and depth (e.g., the pattern reaches the foreground object before the background). **Right:** We show transient light transport under multiple colored point lights placed at different positions.

7.3 Time-Resolved Relighting

Recovering plenoptic light transport enables physically consistent relighting beyond steady-state appearance. Unlike conventional relighting methods that operate on time-integrated radiance, time-resolved relighting allows us to visualize when and how individual light paths contribute to novel illuminations.

Recent work [Malik et al. 2025] demonstrates this impressive effect by refitting the neural network to a new lighting pattern after being trained on lidar view transient light propagation. In this formulation, relighting is achieved through representation refitting rather than explicit manipulation of illumination parameters.

Our framework reconstructs the scene not as a neural representation, but as a set of physically meaningful parameters. This decomposition explicitly enables simulating how light propagates through the captured scene under completely novel illumination conditions. By replacing the original laser source with a new light source, solving the forward transient rendering equation (Eq. (2)), and finally applying the neural compensation, we can synthesize transient videos that visualize the interaction of new illumination within the scene. We demonstrate this capability in Fig. 16. In *Relit Image* column, we project a virtual light source containing a colored pattern (the SIGGRAPH logo) onto the scenes. Our method generates a realistic temporal sequence of light transport inside the scene. As shown in the *Selected Frames* columns, we can observe the structured wavefront of the projected pattern propagating across the scenes. For instance, in the *Bunny Apple Box* scene (top row), the projection hits the front-facing bunny at $t \approx 9.904$ ns before propagating to the background apple and wall at $t \approx 10.784$ ns. Similarly, in the *Metal Sphere* scene (bottom row), the metallic reflection of the projected pattern is accurately synthesized on the sphere's surface. The right two columns of Fig. 16 further illustrates relighting with spatially distinct red, blue and green point lights. Here, we place virtual point sources with different colors at different spatial locations. The resulting transient frames capture the complex interplay of colored wavefronts, dynamic shadows, and indirect light transport. These results validate that our framework recovers a physically robust light transport of the scene under various illuminations.

8 CONCLUSION

We introduce GenPIE, a time-resolved imaging framework that fundamentally rethinks the imaging of plenoptic light transport. By bridging the gap between physically grounded inverse transient rendering and data-driven generative priors, we demonstrate that a high-fidelity and physically grounded reconstruction of geometry and materials is achievable from sparse sampling of the plenoptic light transport.

Our key insight lies in the symbiosis of two distinct paradigms: while foundation models provide the semantic intuition to resolve geometric ambiguities in unobserved regions, the rigorous laws of plenoptic light transport ensure that the final reconstruction remains faithful to the objective physical reality. Crucially, we demonstrate a wide range of applications that were previously considered unattainable without the high-fidelity recovery of the underlying plenoptic light transport.

Limitations and future work. Despite its advantages, GenPIE has several limitations. Differentiable transient path tracing remains

computationally expensive for large-scale scenes with complex geometry, detailed textures, and rich multi-bounce interactions. While foundation models provide strong geometric and semantic priors, reconstruction quality may degrade for objects that fall outside the training distribution. In addition, our pipeline relies on accurate camera calibration; in the current framework, metric scale is recovered by manually measuring distances in the reconstructed point cloud produced by a feed-forward 3D vision model (VGGT) and aligning them with known real-world dimensions, which can introduce inaccuracies. This limitation could be alleviated by fine-tuning VGGT with depth supervision. We believe that future advances in transient sensing hardware and more efficient differentiable rendering algorithms will further broaden the applicability and scalability of our approach.

Ultimately, our current reliance on general-purpose vision models represents a transitional phase; We believe that future foundation models will move beyond priors learned solely from 2D projected imagery and text-conditioned image synthesis. Instead, by training directly on large-scale light transport datasets, such models can internalize the underlying physical principle, i.e., the Rendering Equation, rather than merely approximating its visual outcomes. This shift is expected to enable a deeper and more principled understanding of scene structure, light–matter interaction, and temporal propagation, thereby endowing foundation models with intrinsic physical reasoning capabilities and narrowing the long-standing gap between digital generation and physical reality.

Acknowledgments

We would like to thank the anonymous reviewers for their invaluable suggestions. We also thank Dongyu Du for discussions on scattering imaging experiments, and Guan Huang and Ruiqian Li for their help with hardware system construction. This work was supported in part by the National Natural Science Foundation of China under Grants W2431046 and 61977047, National Key R&D Program of China 2025YFA1309603, Central Guided Local Science and Technology Foundation of China YDZX20253100001001, and by MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence.

References

Nils Abramson. 1978. Light-in-flight recording by holography. *Optical Letters* 3, 4 (1978), 121–123.

Edward H. Adelson and James R. Bergen. 1991. *The Plenoptic Function and the Elements of Early Vision.* MIT Press, Cambridge, MA, USA.

Byeongjoo Ahn, Akshat Dave, Ashok Veeraraghavan, Ioannis Gkioulekas, and Aswin C Sankaranarayanan. 2019. Convolutional approximations to the general non-line-of-sight imaging operator. In *ICCV*.

Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. 2022. Learning Neural Light Fields with Ray-Space Embedding Networks. In *CVPR*.

Seung-Hwan Baek and Felix Heide. 2021. Polarimetric spatio-temporal light transport probing. *ACM Trans. Graph.* 40, 6, Article 212 (Dec. 2021), 18 pages.

Jonathan T. Barron and Jitendra Malik. 2014. Shape, Illumination, and Reflectance from Shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 8 (2014), 1670–1687.

Brent Burley. 2012. Physically Based Shading at Disney. *SIGGRAPH 2012 Course Notes* (2012). https://www.disneyanimation.com/publications/physically-based-shading-at-disney/ Accessed: 2023-11-15.

Brent Burley. 2015. Extending the Disney BRDF to a BSDF with Integrated Subsurface Scattering. *SIGGRAPH 2015 Course Notes* (2015). https://blog.selfshadow.com/publications/s2015-shading-course/#course_content Accessed: 2023-11-15.

Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. 2015. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics express* 23, 16 (2015), 20997–21011.

Wenzheng Chen, Fangyin Wei, Kiriakos N Kutulakos, Szymon Rusinkiewicz, and Felix Heide. 2020. Learned feature embeddings for non-line-of-sight imaging and recognition. *ACM Transactions on Graphics* 39, 6 (2020), 1–18.

Kristin Dana, Bram Van Ginneken, Shree K. Nayar, and Jan J. Koenderink. 1999. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics* 18, 1 (1999), 1–34.

Paul Debevec. 2012. The Light Stages and Their Applications to Photoreal Digital Actors. In *SIGGRAPH Asia*.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *SIGGRAPH*.

Dongyu Du, Xin Jin, Rujia Deng, Jinshi Kang, Hongkun Cao, Yihui Fan, Zhiheng Li, Haoqian Wang, Xiangyang Ji, and Jingyan Song. 2022. A Boundary Migration Model for Imaging within Volumetric Scattering Media. *Nature Communications* 13, 1 (June 2022), 3234.

Faro Technologies Inc. 2012. Measuring Arms. http://www.faro.com.

Xiaohua Feng, Yayao Ma, and Liang Gao. 2022. Compact light field photography towards versatile three-dimensional vision. *Nature Communications* 13 (2022), 333: 1–10.

Yuki Fujimura, Takahiro Kushida, Takuya Funatomi, and Yasuhiro Mukaigawa. 2023. NLOS-NeuS: Non-line-of-sight Neural Implicit Surface. In *CVPR*.

Yasutaka Furukawa and Jean Ponce. 2010. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 8 (2010), 1362–1376.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xuanda Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Xin Xia, Xuefeng Xiao, Zhonghua Zhai, Xinyu Zhang, Qi Zhang, Yuwei Zhang, Shijia Zhao, Jianchao Yang, and Weilin Huang. 2025. Seedream 3.0 Technical Report. arXiv:2504.11346 [cs.CV] https://arxiv.org/abs/2504.11346

Genevieve Gariepy, Nikola Krstajić, Robert Henderson, Chunyong Li, Robert R. Thomson, Gerald S. Buller, Barmak Heshmat, Ramesh Raskar, Jonathan Leach, and Daniele Faccio. 2015. Single-photon sensitive light-in-fight imaging. *Nature Communications* 6, 1 (2015), 6021.

Wenhang Ge, Jiantao Lin, Guibao Shen, Jiawei Feng, Tao Hu, Xinli Xu, and Ying-Cong Chen. 2025. PRM: Photometric Stereo based Large Reconstruction Model. In *ICCV*.

Ioannis Gkioulekas, Anat Levin, Fredo Durand, and Todd Zickler. 2015. Femto-photography: capturing and visualizing the propagation of light. *ACM Transactions on Graphics* 34, 4 (2015), 37.

Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. 1996. The lumigraph. In *SIGGRAPH*.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adrsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izdi. 2019. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics* 38, 6 (2019), 217: 1–19.

Aaron Hertzmann and Steven M. Seitz. 2005. Example-Based Photometric Stereo: Shape Reconstruction with General, Varying BRDFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 8 (2005), 1254–1264.

David S. Immel, Michael F. Cohen, and Donald P. Greenberg. 1986. A radiosity method for non-diffuse environments. *Annual Conference on Computer Graphics and Interactive Techniques* 20, 4 (Aug. 1986), 133–142.

Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. 2022. *Mitsuba 3 renderer.* https://mitsuba-renderer.org.

James T. Kajiya. 1986. The rendering equation. In *Annual Conference on Computer Graphics and Interactive Techniques*. ACM SIGGRAPH.

Sing Bing Kang, Yin Li, Xin Tong, and Heung-Yeung Shum. 2006. *Image-Based Rendering, Foundation and Trends in Computer Graphics and Vision.* Vol. 2. NOW, MA, USA.

Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. 2006. Poisson surface reconstruction. In *Eurographics Symposium on Geometry Processing*. 61–70.

Ahmed Kirmani, Dheera Venkatraman, Dongeek Shin, Andrea Colaco, Franco N. C. Wond, Jeffrey H. Shapiro, and Vivek K Goyal. 2014. First-Photon Imaging. *Science* 343 (2014), 58–61.

Tzofi Klinghoffer, Xiaoyu Xiang, Siddharth Somasundaram, Yuchen Fan, Christian Richardt, Ramesh Raskar, and Rakesh Ranjan. 2024. PlatoNeRF: 3D Reconstruction in Plato's Cave via Single-View Two-Bounce Lidar. In *CVPR*.

Jeremy Klotz and Shree K. Nayar. 2024. Minimalist Vision with Freeform Pixels. In *ECCV*.

Alankar Kotwal, Anat Levin, and Ioannis Gkioulekas. 2023. Passive Micron-scale Time-of-Flight with Sunlight Interferometry. In *CVPR*. IEEE.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. 2024. Grounding Image Matching in 3D with MASt3R.

Marc Levoy and Pat Hanrahan. 1996. Light field rendering. In *SIGGRAPH*.

Yue Li, Yueyi Zhang, Juntian Ye, Feihu Xu, and Zhiwei Xiong. 2023. Deep Non-line-of-sight Imaging from Under-scanning Measurements. In *NeurIPS*.

Gao Liang, Jinyang Liang, Chiye Li, and Lihong V. Wang. 2014. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature* 516 (2014), 74–77.

David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. 2018. Towards transient imaging at interactive rates with singlephoton detectors. In *ICCP*. IEEE.

David B Lindell and Gordon Wetzstein. 2020. Three-dimensional imaging through scattering media based on confocal diffuse tomography. *Nature Communications* 11, 4517 (2020), 1–13.

David B Lindell, Gordon Wetzstein, and Matthew O'Toole. 2019. Wave-based non-line-of-sight imaging using fast fk migration. *ACM Transactions on Graphics* 38, 4 (2019), 1–13.

Xiaochun Liu, Sebastian Bauer, and Andreas Velten. 2020. Phasor filed diffraction based reconstruction for fast non-line-of-sight imaging systems. *Nature communications* 11 (2020), 1645.

Xiaochun Liu, Ibón Guillén, Marco La Manna, Ji Hyun Nam, Syed Azer Reza, Toan Huu Le, Adrian Jarabo, Diego Gutierrez, and Andreas Velten. 2019. Non-line-of-sight imaging using phasor-field virtual wave optics. *Nature* 572, 7771 (2019), 620–623.

Xintong Liu, Jianyu Wang, Leping Xiao, Zuoqiang Shi, Xing Fu, and Lingyun Qiu. 2023. Non-line-of-sight imaging with arbitrary illumination and detection pattern. *Nature Communications* 14, 1 (2023), 3230.

Weihan Luo, Anagh Malik, and David B. Lindell. 2025. Transientangelo: Few-Viewpoint Surface Reconstruction Using Single-Photon Lidar. In *WACV*. 8723–8733.

Anagh Malik, Benjamin Attal, Andrew Xie, Matthew O'Toole, and David Lindell. 2025. Neural Inverse Rendering from Propagating Light. In *CVPR*.

Anagh Malik, Noah Juravsky, Ryan Po, Gordon Wetzstein, Kiriakos N Kutulakos, and David B Lindell. 2024. Flying with photons: Rendering novel views of propagating light. In *ECCV*. Springer, 333–351.

Anagh Malik, Parsa Mirdehghan, Sotiris Nousias, Kiriakos N. Kutulakos, and David B. Lindell. 2023a. Transient Neural Radiance Fields for Lidar View Synthesis and 3D Reconstruction. *NeurIPS*.

Anagh Malik, Parsa Mirdehghan, Sotiris Nousias, Kiriakos N. Kutulakos, and David B. Lindell. 2023b. Transient Neural Radiance Fields for Lidar View Synthesis and 3D Reconstruction. *NeurIPS*.

Leonard McMillan and Gary Bishop. 1995. Plenoptic modeling: an image-based rendering system *(SIGGRAPH '95)*. Association for Computing Machinery, New York, NY, USA, 39–46.

Leonard McMillan and Steven Gortler. 1999. Image-based rendering: A new interface between computer vision and computer graphics. *ACM Transactions on Graphics* 33, 4 (1999), 61–64.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *CVPR*.

Mark Meyer, Mathieu Desbrun, Peter Schröder, and Alan H. Barr. 2003. Discrete Differential-Geometry Operators for Triangulated 2-Manifolds. In *Visualization and Mathematics III*, Hans-Christian Hege and Konrad Polthier (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 35–57.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:2003.08934 [cs]

Ben Mildhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Fangzhou Mu, Sicheng Mo, Jiayong Peng, Xiaochun Liu, Ji Hyum Nam, and Siddeshar Raghavan. 2025. Physics to the Rescue: Deep Non-Line-of-Sight Reconstruction for High-Speed Imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 8 (2025), 6164–6158.

Fangzhou Mu, Carter Sifferman, Sacha Hungerman, Yiquan Li, Mark Han, Michael Gleicher, Mohit Gupta, and Yin Li. 2024. Towards 3D Vision with Low-Cost Single-Photon Cameras. In *CVPR*. IEEE.

Ren Ng, Marc Levoy, Mathieu Bredif, Gene Ducal, Mark Horowitz, and Pat Hanrahan. 2005. Light Field Photography with a Hand-held Plenoptic Camera. In *Stanford Tech Report CTSR*.

Baptiste Nicolet, Alec Jacobson, and Wenzel Jakob. 2021. Large steps in inverse rendering of geometry. *ACM Transactions on Graphics* (2021).

OpenAI. 2025. GPT Image 1.5. https://platform.openai.com/docs/models/gpt-image-1.5.

Matthew O'Toole, Felix Heide, Lei Xiao, Matthias B. Hullin, Wolfgang Heidrich, and Kiriakos N. Kutulakos. 2014. Temporal Frequency Probing for 5D Transient Analysis of Global Light Transport. *ACM Transactions on Graphics* 33, 4 (2014), 87.

Matthew O'Toole, Felix Heide, David B. Lindell, Kai Zang, Steven Diamond, and Gordon Wetzstein. 2017. Reconstructing Transient Images from Single-Photon Sensors. In *CVPR*. IEEE.

Matthew O'Toole, David B Lindell, and Gordon Wetzstein. 2018. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature* 555, 7696 (2018), 338–341.

Diego Royo, Miguel Crespo, and Jorge Garcia-Pueyo. 2023a. mitransient. https://github.com/diegoroyo/mitransient.

Diego Royo, Jorge Garcia-Pueyo, Miguel Crespo, Óscar Pueyo-Ciutad, Guillermo Enguita, and Diego Bielsa. 2025. mitransient: Transient light transport in Mitsuba 3. arXiv:2510.25660 [cs.GR] https://arxiv.org/abs/2510.25660

Diego Royo, Jorge García, Adolfo Muñoz, and Adrian Jarabo. 2022. Non-line-of-sight transient rendering. *Computers & Graphics* (2022). https://www.sciencedirect.com/science/article/pii/S0097849322001200

Diego Royo, Talha Sultan, Adolfo Muñoz, Khadijeh Masumnia-Bisheh, Eric Brandt, Diego Gutierrez, Andreas Velten, and Julio Marco. 2023b. Virtual Mirrors: Non-Line-of-Sight Imaging Beyond the Third Bounce. *ACM Transactions on Graphics* 42, 4 (2023).

Guy Satat, Matthew Tancik, and Ramesh Raskar. 2018. Towards Photography Through Realistic Fog. In *ICCP*.

Imari Sato, Takahiro Okabe, Yoichi Sato, and Katsushi Ikeuchi. 2003. Appearance sampling for obtaining a set of basis images for variable illumination. In *ICCV*.

Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *CVPR*.

Siyuan Shen, Zi Wang, Ping Liu, Zhengqing Pan, Ruiqian Li, Tian Gao, Shiying Li, and Jingyi Yu. 2021. Non-line-of-sight imaging via neural transient fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 7 (2021), 2257–2268.

Siyuan Shen, Suan Xia, Xingyue Peng, Ziyu Wang, Yingsheng Zhu, Shiying Li, and Jingyi Yu. 2025. HOLI-1-to-3: Transient-enhanced holistic image-to-3d generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 9 (2025), 7206–7217.

Dongeek Shin, Feihu Xu, Dheera Venkatraman, Rudi Lussana, Federica Villa, Franco Zappa, Vivek K Goyal, Franco NC Wong, and Jeffrey H Shapiro. 2016. Photon-efficient imaging with a single-photon camera. *Nature communications* 7, 1 (2016), 12046.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. DINOv3. arXiv:2508.10104 [cs.CV] https://arxiv.org/abs/2508.10104

Adam Smith, James Skorupski, and James Davis. 2008. Transient rendering. *Technic Report UCSC-SOE-08-26* (2008).

Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. 2021. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*.

Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Moungi G Bawendi, and Ramesh Raskar. 2012. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature communications* 3, 1 (2012), 745.

Andreas Velten, Di Wu, Adrian Jarabo, Belen Masia, Christopher Barsi, Chinmaya Joshi, Everett Lawson, Moungi Bawendi, Diego Gutierrez, and Ramesh Raskar. 2013. Femto-photography: capturing and visualizing the propagation of light. *ACM Transactions on Graphics* 32, 4 (2013), 1–8.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. VGGT: Visual Geometry Grounded Transformer. In *CVPR*.

Yuehan Wang, Siyuan Shen, Suan Xia, Ruiqian Li, Xingyue Peng, Yanhua Yu, Shiying Li, and Jingyi Yu. 2023. Neural Reconstruction through Scattering Media with Forward and Backward Losses. In *ICCP*. 1–12.

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

Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. 2005. High performance imaging using large camera arrays. *ACM Transactions on Graphics* 24, 3 (2005), 765–776.

Robert J. Woodham. 1980. Photometric Method for Determining Surface Orientation from Multiple Images. *Optical Engineering* 19, 1 (1980), 139–144.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025. Qwen-Image Technical Report. arXiv:2508.02324 [cs.CV] https://arxiv.org/abs/2508.02324

Di Wu, Andreas Velten, Matthew O'Toole, Belen Masia, Amit Agrawal, Qionghai Dai, and Ramesh Raskar. 2014. Decomposing Global Light Transport Using Time of Flight Imaging. *ICCV* 107 (2014), 123–138.

Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. 2017. Light Field Image Processing: An Overview. *IEEE Journal of Selected Topics in Signal Processing* 11, 7 (2017), 926–954.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. 2025. Native and Compact Structured Latents for 3D Generation. *Tech report* (2025).

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. *arXiv preprint arXiv:2412.01506* (2024).

Shumian Xin, Sotiris Nousias, Kiriakos N Kutulakos, Aswin C Sankaranarayanan, Srinivasa G Narasimhan, and Ioannis Gkioulekas. 2019. A theory of Fermat paths for non-line-of-sight shape reconstruction. In *CVPR*.

Jun-Tian Ye, Xin Huang, Zheng-Ping Li, and Feihu Xu. 2021. Compressed sensing for active non-line-of-sight imaging. *Optics Express* 29, 2 (2021), 1749–1763.

Shinyoung Yi, Donggun Kim, Kiseok Choi, Adrian Jarabo, Diego Gutierrez, and Min H. Kim. 2021. Differentiable transient rendering. *ACM Trans. Graph.* 40, 6, Article 286 (Dec. 2021), 11 pages.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. 2024. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *ACM Transactions on Graphics* 43, 4 (2024), 1–20.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. 586–595.

Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. 1999. Shape from Shading: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 8 (1999), 690–706.

Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. 2021. NeRFactor: neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics* 40, 6 (2021), 237: 1–18.

Hui Zhou, Yuhao He, Lixing You, Sijin Chen, Weijun Zhang, Junjie Wu, Zhen Wang, and Xiaoming Xie. 2015. Few-photon imaging at 1550 nm using a low-timing-jitter superconducting nanowire singlephoton detector. *Optics Express* 23, 11 (2015), 14603–14611.