

# Spatiotemporal Flux Probing for Single-Photon Videography: Supplemental Document

Jerry Yan<sup>1\*</sup>      Matteo Forlivesi<sup>1\*</sup>      Bowen Tan<sup>1</sup>      Andrew Xie<sup>2</sup>  
Siddharth Somasundaram<sup>3</sup>      Sotiris Nousias<sup>1</sup>

<sup>1</sup>Dept. of Computer Science, Purdue University

<sup>2</sup>Dept. of Computer Science, University of Toronto

<sup>3</sup>MIT Media Lab, Massachusetts Institute of Technology

## Contents

|                                                                                                  |           |
|--------------------------------------------------------------------------------------------------|-----------|
| <b>A Preliminaries</b>                                                                           | <b>4</b>  |
| A.1 Mathematical definitions . . . . .                                                           | 4         |
| A.1.1 Spatiotemporal Poisson point process definitions . . . . .                                 | 4         |
| <b>B Spatiotemporal Flux Probing Theory</b>                                                      | <b>6</b>  |
| B.1 Spatiotemporal photon counting process . . . . .                                             | 6         |
| B.2 Spatiotemporal flux probing (proof of Equation 1) . . . . .                                  | 7         |
| B.3 Noise model for probing measurements in Equation 1 . . . . .                                 | 8         |
| B.3.1 Mean and variance of probing measurements . . . . .                                        | 8         |
| B.3.2 Covariance of probing measurements . . . . .                                               | 9         |
| B.4 Noise model for Fourier probing measurements (Equation 2) . . . . .                          | 10        |
| B.4.1 Mean and covariance matrix of Fourier probing measurements (proof of Equation 3) . . . . . | 11        |
| B.4.2 Fourier probing energy (proof of Equation 4) . . . . .                                     | 12        |
| B.5 CFAR frequency detection (proof of Equation 5) . . . . .                                     | 12        |
| <b>C Velocity Detection</b>                                                                      | <b>14</b> |
| C.1 Fourier planar structures induced by linear velocity (proof of Equation 6) . . . . .         | 14        |
| C.2 Parametric velocity detection model . . . . .                                                | 16        |
| C.3 Limitation of the parametric velocity detection model . . . . .                              | 17        |
| C.4 Nonparametric velocity detection model (proof of Equation 8) . . . . .                       | 17        |
| <b>D Extension to Other Asynchronous Sensors</b>                                                 | <b>20</b> |
| D.1 Spike Cameras: Spatiotemporal Probing Derivation . . . . .                                   | 20        |
| D.1.1 Temporal spike probing . . . . .                                                           | 20        |

---

\* Equal contribution: {yan569,mforlive}@purdue.edu

|          |                                                                                         |           |
|----------|-----------------------------------------------------------------------------------------|-----------|
| D.1.2    | Spatiotemporal spike probing . . . . .                                                  | 21        |
| D.1.3    | Fourier probing: null distribution . . . . .                                            | 22        |
| D.1.4    | CFAR detector for spike cameras . . . . .                                               | 22        |
| D.2      | Event Cameras: Spatiotemporal Probing Derivation . . . . .                              | 23        |
| D.2.1    | Temporal event probing . . . . .                                                        | 24        |
| D.2.2    | Spatiotemporal event probing (proof of Equation 9) . . . . .                            | 25        |
| D.2.3    | Simplified noise model for events . . . . .                                             | 26        |
| D.2.4    | Fourier probing: null distribution . . . . .                                            | 26        |
| D.2.5    | CFAR detector for event cameras . . . . .                                               | 28        |
| <b>E</b> | <b>Photon-Location Prediction and Global Flicker: an Intuitive Analysis for bit2bit</b> | <b>30</b> |
| <b>F</b> | <b>Limitations of Per-Pixel Temporal Analysis Under Motion</b>                          | <b>31</b> |
| <b>G</b> | <b>Implementation Details</b>                                                           | <b>33</b> |
| G.1      | Single-photon videography . . . . .                                                     | 33        |
| G.1.1    | Frequency selection and spatiotemporal normalization . . . . .                          | 33        |
| G.1.2    | Selection of the CFAR significance level $\alpha$ . . . . .                             | 33        |
| G.1.3    | Short-time Fourier transform for quanta sensors . . . . .                               | 33        |
| G.1.4    | Windowed CFAR detection and reconstruction . . . . .                                    | 34        |
| G.2      | Velocity-selective videography . . . . .                                                | 34        |
| G.2.1    | Velocity detection hyperparameters . . . . .                                            | 34        |
| G.2.2    | Velocity detection with single-photon sensors . . . . .                                 | 35        |
| G.2.3    | Velocity-selective videography with single-photon sensors . . . . .                     | 35        |
| G.3      | Spike camera videography . . . . .                                                      | 36        |
| G.4      | Event camera videography . . . . .                                                      | 36        |
| G.4.1    | Log-flux estimation from only events . . . . .                                          | 36        |
| G.4.2    | Kalman filter with intensity frames . . . . .                                           | 37        |
| <b>H</b> | <b>Experiments</b>                                                                      | <b>38</b> |
| H.1      | Single-photon videography . . . . .                                                     | 38        |
| H.1.1    | Additional details for Balloon popping with 31 kHz LED (Figure 1 scene) . . . . .       | 39        |
| H.1.2    | Additional details for ultra-wideband videography (Figure 4, row 5 scene) . . . . .     | 39        |
| H.1.3    | Runtime for single-photon videography . . . . .                                         | 39        |
| H.2      | Velocity-selective videography on the Figure 1 scene . . . . .                          | 39        |
| H.3      | Videography with other asynchronous sensors (Figure 5) . . . . .                        | 40        |
| <b>I</b> | <b>Simulations</b>                                                                      | <b>42</b> |
| I.1      | Probing measurement normality . . . . .                                                 | 42        |
| I.1.1    | Central limit theorem validity . . . . .                                                | 42        |
| I.1.2    | Empirical validation . . . . .                                                          | 42        |

|       |                                                             |    |
|-------|-------------------------------------------------------------|----|
| I.2   | CFAR threshold validity . . . . .                           | 42 |
| I.3   | Quantitative reconstruction error . . . . .                 | 43 |
| I.3.1 | Simulating photon arrivals from high-speed videos . . . . . | 43 |
| I.3.2 | Simulated evaluation . . . . .                              | 44 |
| I.4   | Velocity detection . . . . .                                | 46 |
| I.4.1 | Accuracy of the best detection . . . . .                    | 46 |
| I.4.2 | Detection set quality for a single moving object . . . . .  | 47 |
| I.4.3 | Detection accuracy across the velocity range . . . . .      | 48 |
| I.4.4 | Detection accuracy with multiple moving objects . . . . .   | 49 |

## A Preliminaries

**Photon arrivals as a spatiotemporal point process.** While camera sensors physically consist of a discrete array of photodetectors, the photon arrival process can be modeled as a continuous process in space and time, registering the precise location and arrival time of each photon arrival event [12]. Under this framework, over an exposure time  $t_{\text{exp}}$ , photon arrivals are modeled as an inhomogeneous Poisson point process  $\mathcal{P}$  over the spatiotemporal domain  $\Omega = A \times [0, t_{\text{exp}}]$ , where  $A = [0, w] \times [0, h]$  is the spatial area of the detector. Its underlying flux is given by  $\phi$ .

**Spatiotemporal flux.** While flux is normally modeled in photons/second as a temporal function where  $\phi(t) \geq 0$ , this can be extended to the spatiotemporal domain. For a point  $\mathbf{x} = (x, y, t)$ , its instantaneous flux can be written as a function in spacetime, where  $\phi(\mathbf{x}) \geq 0$ . In terms of physical sensor units, this flux can be interpreted as the number of photon arrivals in the units of photons/second/pixel<sup>2</sup>.

Importantly, we restrict the spatiotemporal flux functions that we consider to those that are **nonnegative**, **measurable**, and **bounded**. We also assume that we observe photon arrivals in a **finite** sensor area with finite exposure time. These are critical assumptions that allow spatiotemporal flux probing to hold mathematically.

### A.1 Mathematical definitions

In order to rigorously characterize photon arrivals in space and time, we include definitions of the foundational measure-theoretic and probabilistic structures used throughout this work. These definitions formalize the notion that photons arrive randomly across both space and time, and that our uncertainty about future arrivals can be tracked as observations accumulate.

**Definition 1** ( $\sigma$ -algebra (Section A1.4 in [7])). A  $\sigma$ -algebra  $\mathbf{F}$  on a set  $\Omega$  is a collection of subsets of  $\Omega$  that contains the empty set and is closed under complement and countable unions. The pair  $(\Omega, \mathbf{F})$  is called a measurable space.

In the context of photon arrivals,  $\mathbf{F}$  formalizes all spatiotemporal regions in  $\Omega$  that are well-defined and observable—for example, the region corresponding to a given pixel area during a given time interval. Together, the elements of  $\mathbf{F}$  represent every possible collection of photon arrival events that a sensor could observe over its exposure.

**Definition 2** (Measurability (Section A1.4 in [7])). Let  $(\Omega, \mathbf{F})$  be a measurable space.

- (i) A set  $W \subseteq \Omega$  is measurable if  $W \in \mathbf{F}$ .
- (ii) A function  $f : \Omega \rightarrow \mathbb{R}$  is measurable if for any measurable set  $B \subseteq \mathbb{R}$ , the set  $f^{-1}(B) = \{\mathbf{x} \in \Omega : f(\mathbf{x}) \in B\}$  is measurable.

Measurability of flux function  $\phi$  ensures that  $\int_W \phi(\mathbf{x}) d\mathbf{x}$  is well-defined for any measurable  $W \subseteq \Omega$ , so that the expected number of photon arrivals in any observable region can be computed via integration against  $\phi$ .

#### A.1.1 Spatiotemporal Poisson point process definitions

**Definition 3** (Spatiotemporal filtration). A spatiotemporal filtration is a family  $\{\mathbf{F}_t\}_{t \in [0, t_{\text{exp}}]}$  of  $\sigma$ -algebras on  $\Omega$ , indexed by time  $t$ , such that  $\mathbf{F}_s \subseteq \mathbf{F}_t$  for all  $s \leq t$ . Each  $\mathbf{F}_t$  encodes all observable information accumulated up to time  $t$ , across the full sensor domain  $A \subset \mathbb{R}^2$ .

Intuitively, the filtration represents the growing history of photon arrivals on the sensor: as the exposure progresses, more photons have landed and  $\mathbf{F}_t$  grows to reflect all spatial arrival patterns observed up to time  $t$ .

**Definition 4** (Spatiotemporal adapted process (Definition 3.2.14 in [19])). A stochastic process  $\{X_t\}_{t \in [0, t_{\text{exp}}]}$ , where each  $X_t$  is a spatially-indexed random field over  $A$ , is adapted to the filtration  $\{\mathbf{F}_t\}$  if for each  $t$ , the random field  $X_t(\cdot)$  is  $\mathbf{F}_t$ -measurable. That is, the value of  $X_t$  at any spatial location  $\mathbf{r} \in A$  is fully determined by the information available up to time  $t$ .

For photon arrivals, adaptedness means that our description of the arrival process at any spatial location on the sensor at time  $t$  depends only on what has actually been observed up to that moment—not on future photon arrivals.

**Definition 5** (Spatiotemporal  $\mathbf{F}$ -martingale). Let  $\{\mathbf{F}_t\}_{t \in [0, t_{\text{exp}}]}$  be a filtration. A process  $M(t) = M(A \times [0, t])$ , aggregated over sensor domain  $A \subset \mathbb{R}^2$ , is an  $\mathbf{F}$ -martingale if it is  $\mathbf{F}$ -adapted and:

- (i)  $\mathbb{E}[|M(t)|] < \infty$  for all  $t \in [0, t_{\text{exp}}]$ ,
- (ii)  $\mathbb{E}[M(t) \mid \mathbf{F}_s] = M(s)$  a.s. for all  $s \leq t$ .

Intuitively, a spatiotemporal martingale describes a quantity defined across the sensor whose expected future value at any pixel location, given everything observed so far, is simply its current value. In the context of photon arrivals, a martingale arises when we correct the raw arrival counts by their expected flux contribution, leaving a residual process that carries no predictable trend forward in time.

**Definition 6** (Inhomogeneous Poisson point process). A spatiotemporal point process  $\mathcal{P}$  on  $\Omega = A \times [0, t_{\text{exp}}] \subset \mathbb{R}^2 \times \mathbb{R}^+$  is an inhomogeneous Poisson point process with flux function  $\phi : \Omega \rightarrow \mathbb{R}^+$  if:

- (i) For any measurable  $W \subseteq \Omega$ , the count  $\mathcal{P}(W) \sim \text{Poisson}(\Phi(W))$ , where the intensity measure is

$$\Phi(W) = \int_W \phi(\mathbf{x}) d\mathbf{x} \tag{10}$$

- (ii) For disjoint  $W_1, \dots, W_k \subseteq \Omega$ , the counts  $\mathcal{P}(W_1), \dots, \mathcal{P}(W_k)$  are independent.

For simplicity, we denote the total number of observed photon arrivals as  $|\mathcal{P}| = \mathcal{P}(\Omega)$ .

This models photons arriving at a sensor as random events in space and time, where the local flux  $\phi(\mathbf{r}, t)$  governs how frequently photons tend to land near location  $\mathbf{r}$  at time  $t$ . Brighter regions of a scene produce higher flux and therefore more photon arrivals, and arrivals in nonoverlapping regions or time intervals are statistically independent of one another.

## B Spatiotemporal Flux Probing Theory

### B.1 Spatiotemporal photon counting process

We start by showing the martingale decomposition of Poisson point process  $\mathcal{P}$ .

**Proposition 1** (Poisson point process decomposition). Let  $\mathcal{P}(\Omega) = \mathcal{P}(A \times [0, t_{\text{exp}}])$  be the inhomogeneous Poisson point process with flux function  $\phi(\mathbf{x})$ , and let  $\mathbf{F}$  be a filtration such that  $\mathcal{P}(\Omega)$  is  $\mathbf{F}$ -adapted. Then for any measurable subset  $W = A \times [0, t] \subseteq \Omega$  where  $0 < t \leq t_{\text{exp}}$ ,

$$M(t) = \mathcal{P}(W) - \int_W \phi(\mathbf{x}) d\mathbf{x}, \quad (11)$$

where  $M(t) = M(A \times [0, t])$ .

In the special case that  $\mathcal{P}$  is a 1D inhomogeneous Poisson point process, such that it is only defined on  $[0, t_{\text{exp}}]$ , this decomposition can instead be written in terms of counting process  $N(t) = \mathcal{P}([0, t])$  [41]:

$$M(t) = N(t) - \int_0^t \phi(u) du. \quad (12)$$

*Proof Sketch of Proposition 1.* The proof proceeds in two steps:

**Step 1:** We show the Poisson process  $\mathcal{P}$  can be decomposed in terms of its flux  $\phi(\mathbf{x})$  and zero-mean noise  $M$ . To do this, we make use of the following lemma:

**Lemma 1** (Poisson compensated integral (Equation 12.4 in [24])). Define Poisson point process  $\mathcal{P}$  with flux function  $\phi$  on the domain  $\Omega = A \times [0, t_{\text{exp}}]$ . For any deterministic function  $f$  such that  $\int_{\Omega} |f(\mathbf{x})\phi(\mathbf{x})| d\mathbf{x} < \infty$ , then  $\int_W |f(\mathbf{x})\phi(\mathbf{x})| d\mathbf{x} < \infty$  on any measurable subset  $W = A \times [0, t] \subseteq \Omega$ . The compensated integral of  $f$  with respect to  $\mathcal{P}$  is defined by

$$\int_W f(\mathbf{x}) dM(\mathbf{x}) = \int_W f(\mathbf{x}) d\mathcal{P}(\mathbf{x}) - \int_W f(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x}. \quad (13)$$

such that  $\mathbb{E} \left[ \int_W f(\mathbf{x}) dM(\mathbf{x}) \right] = 0$ .

**Step 2:** We show that the stochastic process  $M$  satisfies the conditions for an  $\mathbf{F}$ -martingale, such that  $\mathbb{E}[|M(t)|] < \infty$  and  $\mathbb{E}[M(t) | \mathbf{F}_s] = M(s)$ .  $\square$

*Proof of Proposition 1. Step 1:* Let  $f(\mathbf{x}) = 1$ . Since  $\phi$  is measurable, nonnegative, and bounded on a finite domain, then  $\int_W |\phi(\mathbf{x})| d\mathbf{x} = \int_W \phi(\mathbf{x}) d\mathbf{x} < \infty$ . Then we can write its noise term as

$$\int_W dM(\mathbf{x}) = M(W) = \mathcal{P}(W) - \int_W \phi(\mathbf{x}) d\mathbf{x}, \quad \mathbb{E}[M(W)] = 0. \quad (14)$$

We can parametrize  $M(W)$  by time such that  $M(t) = M(A \times [0, t])$ .

**Step 2:** First, we show that  $\mathbb{E}[|M(t)|] < \infty$ . From Eq. (14) and the triangle inequality,

$$\mathbb{E}[|M(t)|] = \mathbb{E} \left[ \left| \mathcal{P}(W) - \int_W \phi(\mathbf{x}) d\mathbf{x} \right| \right] \quad (15a)$$

$$\leq \underbrace{\mathbb{E}[|\mathcal{P}(W)|]}_{\mathcal{P}(W) \geq 0} + \underbrace{\mathbb{E} \left[ \left| \int_W \phi(\mathbf{x}) d\mathbf{x} \right| \right]}_{\phi(\mathbf{x}) \geq 0 \text{ and deterministic}} \quad (15b)$$

$$= \underbrace{\mathbb{E}[\mathcal{P}(W)]}_{=\int_W \phi(\mathbf{x}) d\mathbf{x}} + \int_W \phi(\mathbf{x}) d\mathbf{x} \quad (15c)$$

$$= 2 \int_W \phi(\mathbf{x}) d\mathbf{x} < \infty \quad (15d)$$

Next, we show  $\mathbb{E}[M(t) \mid \mathbf{F}_s] = M(s)$  for all  $s \leq t$ . Write

$$\mathbb{E}[M(t) \mid \mathbf{F}_s] = \mathbb{E} \left[ \mathcal{P}(A \times [0, t]) - \underbrace{\int_{A \times [0, t]} \phi(\mathbf{x}) d\mathbf{x}}_{\text{deterministic}} \mid \mathbf{F}_s \right] \quad (16a)$$

$$= \mathbb{E}[\mathcal{P}(A \times [0, t]) \mid \mathbf{F}_s] - \int_{A \times [0, t]} \phi(\mathbf{x}) d\mathbf{x}, \quad (16b)$$

We decompose the count into the observed past and unobserved future,

$$\mathcal{P}(A \times [0, t]) = \mathcal{P}(A \times [0, s]) + \mathcal{P}(A \times (s, t]), \quad (17)$$

so that

$$\mathbb{E}[\mathcal{P}(A \times [0, t]) \mid \mathbf{F}_s] = \mathcal{P}(A \times [0, s]) + \mathbb{E}[\mathcal{P}(A \times (s, t]) \mid \mathbf{F}_s] \quad (18a)$$

$$= \mathcal{P}(A \times [0, s]) + \int_{A \times (s, t]} \phi(\mathbf{x}) d\mathbf{x}, \quad (18b)$$

where  $\mathcal{P}(A \times [0, s])$  is  $\mathbf{F}_s$ -measurable and the increment  $\mathcal{P}(A \times (s, t])$  is independent of  $\mathbf{F}_s$  by the independent increments property of the Poisson process, with conditional expectation equal to its intensity integral. Substituting back,

$$\mathbb{E}[M(t) \mid \mathbf{F}_s] = \mathcal{P}(A \times [0, s]) + \int_{A \times (s, t]} \phi(\mathbf{x}) d\mathbf{x} - \int_{A \times [0, t]} \phi(\mathbf{x}) d\mathbf{x} \quad (19a)$$

$$= \mathcal{P}(A \times [0, s]) - \int_{A \times [0, s]} \phi(\mathbf{x}) d\mathbf{x} \quad (19b)$$

$$= M(s). \quad (19c)$$

Since  $\mathbb{E}[|M(t)|] < \infty$ ,  $M(t)$  is  $\mathbf{F}_t$ -measurable by construction, and  $\mathbb{E}[M(t) \mid \mathbf{F}_s] = M(s)$  for all  $s \leq t$ , we conclude that  $M(t)$  is an  $\mathbf{F}$ -martingale.  $\square$

## B.2 Spatiotemporal flux probing (proof of Equation 1)

Using the martingale decomposition in Proposition 1, we can prove the decomposition of the spatiotemporal flux probing measurements in Equation 1 in the main paper. We begin by restating a rigorous version of Equation 1:

**Equation 1.** Define spatiotemporal Poisson point process  $\mathcal{P}$  over the spatiotemporal volume  $[0, w] \times [0, h] \times [0, t_{\text{exp}}]$  with underlying flux function  $\phi(x, y, t)$ . Given a bounded, deterministic, and measurable probing function  $p(x, y, t)$ , the following

decomposition for probing measurements holds:

$$\underbrace{\sum_{(x_i, y_i, t_i) \in \mathcal{P}} p(x_i, y_i, t_i)}_{\text{probing measurements}} = \underbrace{\int_0^{t_{\text{exp}}} \int_0^h \int_0^w p(x, y, t) \phi(x, y, t) dx dy dt}_{\text{flux probing integral } \langle p, \phi \rangle} + \underbrace{M_p(t_{\text{exp}})}_{\text{martingale noise}}. \quad (1)$$

If  $\mathcal{P}$  is a 1D Poisson point process, this becomes identical to the flux probing equation in [41]:

$$\sum_{t_i \in \mathcal{P}} p(t_i) = \int_0^{t_{\text{exp}}} p(t) \phi(t) dt + M_p(t_{\text{exp}}). \quad (20)$$

We restate Equation 1 as a more succinct integral where  $\Omega = [0, w] \times [0, h] \times [0, t_{\text{exp}}]$ , which will be the notation used for the rest of the supplement:

$$\sum_{\mathbf{x}_i \in \mathcal{P}} p(\mathbf{x}_i) = \int_{\Omega} p(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x} + M_p(t_{\text{exp}}). \quad (21)$$

*Proof Sketch of Equation 1.* The proof proceeds in two steps:

**Step 1:** We substitute the function  $f$  for our probing function  $p$  in Lemma 1.

**Step 2:** We derive the flux probing measurements using the following lemma:

**Lemma 2** (Proposition 8.3.2.1.i in [19]). Let  $p(\mathbf{x})$  be a bounded and deterministic function. Then the process

$$M_p(t) = \int_{A \times [0, t]} p(\mathbf{x}) dM(\mathbf{x}) \quad (22)$$

is a **F**-martingale. The cited result establishes this for 1D processes; the spatiotemporal case follows by fixing  $A$  and applying the result to the marginal process  $t \mapsto M(A \times [0, t])$ . □

*Proof of Equation 1. Step 1:* Let  $f(\mathbf{x}) = p(\mathbf{x})$  in Lemma 1. We obtain:

$$\int_{\Omega} p(\mathbf{x}) d\mathcal{P}(\mathbf{x}) = \int_{\Omega} p(\mathbf{x}) dM(\mathbf{x}) + \int_{\Omega} p(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x}. \quad (23)$$

**Step 2:** The left-hand term equals  $\sum_{\mathbf{x} \in \mathcal{P}} p(\mathbf{x})$  by definition of integration against a counting measure. The martingale term

$$t \mapsto M_p(t) = \int_{A \times [0, t]} p(\mathbf{x}) dM(\mathbf{x}) \quad (24)$$

is a **F**-martingale by Lemma 2. Setting  $t = t_{\text{exp}}$  and substituting  $M_p(t_{\text{exp}})$  gives us Equation 1. □

## B.3 Noise model for probing measurements in Equation 1

From this section forward, we denote the probing measurements as  $\mathcal{E}_p \triangleq \sum_{\mathbf{x} \in \mathcal{P}} p(\mathbf{x})$  and the probing flux integral as  $\langle p, \phi \rangle = \int_{\Omega} p(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x}$  for brevity.

### B.3.1 Mean and variance of probing measurements

When describing the noise model for the flux probing measurements in the main paper, we claim the following:

**Proposition 2.** The probing measurements  $\mathcal{E}_p$  are unbiased with mean  $\langle p, \phi \rangle$ , variance  $\langle p^2, \phi \rangle$ , and approximately follow a normal distribution.

*Proof Sketch of Proposition 2. Step 1:* We show the mean of  $\mathcal{E}_p$  is  $\langle p, \phi \rangle$ , its variance is  $\langle p^2, \phi \rangle$ , and that it is unbiased via Campbell's theorem:

**Lemma 3** (Campbell's theorem (Section 3.2 in [23])). For a Poisson point process  $\mathcal{P}$  on  $\Omega$  with flux intensity  $\phi$  and intensity measure  $d\Phi(\mathbf{x}) = \phi(\mathbf{x}) d\mathbf{x}$ , and a measurable function  $p : \Omega \rightarrow \mathbb{R}$ :

(i) If  $\int_{\Omega} |p(\mathbf{x})| \phi(\mathbf{x}) d\mathbf{x} < \infty$ , then

$$\mathbb{E} \left[ \sum_{\mathbf{x} \in \mathcal{P}} p(\mathbf{x}) \right] = \int_{\Omega} p(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x}. \quad (25)$$

(ii) If  $\int_{\Omega} p(\mathbf{x})^2 \phi(\mathbf{x}) d\mathbf{x} < \infty$ , then

$$\text{Var} \left( \sum_{\mathbf{x} \in \mathcal{P}} p(\mathbf{x}) \right) = \int_{\Omega} p(\mathbf{x})^2 \phi(\mathbf{x}) d\mathbf{x}. \quad (26)$$

Since we choose probing functions  $p$  such that they are bounded and measurable, and evaluate them on a bounded domain, these integrals naturally converge.

**Step 2:** We show  $\mathcal{E}_p$  is approximately normal via the central limit theorem. □

*Proof of Proposition 2. Step 1:* The result follows directly from Campbell's theorem:

$$\mathbb{E}[\mathcal{E}_p] = \langle p, \phi \rangle, \quad (27)$$

$$\text{Var}(\mathcal{E}_p) = \langle p^2, \phi \rangle. \quad (28)$$

Naturally, this means our estimator  $\mathcal{E}_p$  is unbiased.

**Step 2:** The probing measurements are a sum of independent identically-distributed random variables, where each photon observation location is a random variable. This approximates a normal distribution under the central limit theorem [33]. We verify this empirically via simulation in Section I.1. □

### B.3.2 Covariance of probing measurements

**Proposition 3.** For probing functions  $p_1, p_2$ , their probing measurement covariance is:

$$\text{Cov}(\mathcal{E}_{p_1}, \mathcal{E}_{p_2}) = \langle p_1 p_2, \phi \rangle \quad (29)$$

*Proof Sketch of Proposition 3.* We introduce the concept of second factorial moment measures for a more generalizable proof that is more succinct and does not require properties of martingales.

**Definition 7** (Second factorial moment measure (Definition 4.9 in [24])). For a general point process  $\mathcal{P}$  and nonnegative measurable function  $f$ , its second factorial moment measure is defined as follows:

$$\mathbb{E} \left[ \sum_{\mathbf{x}, \mathbf{x}' \in \mathcal{P}, \mathbf{x} \neq \mathbf{x}'} f(\mathbf{x}, \mathbf{x}') \right] = \int_{\Omega \times \Omega} f(\mathbf{x}, \mathbf{x}') M^{(2)}(d\mathbf{x}, d\mathbf{x}'). \quad (30)$$

Intuitively, the second factorial moment measure informs us, on average, how many distinct pairs of photons appear inside  $\mathbf{x}$ . For general point processes, this can tell us whether points tend to cluster or spread out from each other. However, since photon arrivals are independent, knowing one photon arrival does not inform us where other photons arrive. This naturally leads to the following lemma:

**Lemma 4** (Corollary 4.10 in [24]). For a Poisson point process with flux function  $\phi$ , its 2nd factorial moment measure simplifies to the following:

$$\int_{\Omega \times \Omega} f(\mathbf{x}, \mathbf{x}') M^{(2)}(d\mathbf{x}, d\mathbf{x}') = \int_{\Omega \times \Omega} f(\mathbf{x}, \mathbf{x}') \phi(\mathbf{x}) \phi(\mathbf{x}') d\mathbf{x} d\mathbf{x}'. \quad (31)$$

□

*Proof of Proposition 3.* The covariance by definition is

$$\text{Cov}(\mathcal{E}_{p_1}, \mathcal{E}_{p_2}) = \mathbb{E}[\mathcal{E}_{p_1} \mathcal{E}_{p_2}] - \mathbb{E}[\mathcal{E}_{p_1}] \mathbb{E}[\mathcal{E}_{p_2}]. \quad (32)$$

We first evaluate  $\mathbb{E}[\mathcal{E}_{p_1} \mathcal{E}_{p_2}]$ :

$$\mathbb{E}[\mathcal{E}_{p_1} \mathcal{E}_{p_2}] = \mathbb{E}\left[\sum_{\mathbf{x} \in \mathcal{P}} p_1(\mathbf{x}) p_2(\mathbf{x})\right] + \mathbb{E}\left[\sum_{\mathbf{x}, \mathbf{x}' \in \mathcal{P}, \mathbf{x} \neq \mathbf{x}'} p_1(\mathbf{x}) p_2(\mathbf{x}')\right] \quad (33)$$

Letting  $f(\mathbf{x}) = p_1(\mathbf{x}) p_2(\mathbf{x})$ , we have from Campbell's theorem (Lemma 3) that

$$\mathbb{E}\left[\sum_{\mathbf{x} \in \mathcal{P}} p_1(\mathbf{x}) p_2(\mathbf{x})\right] = \int_{\Omega} p_1(\mathbf{x}) p_2(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x} = \langle p_1 p_2, \phi \rangle. \quad (34)$$

Then from Definition 7 and Lemma 4, the second term can be written as a second factorial moment measure, and simplified as follows:

$$\mathbb{E}\left[\sum_{\mathbf{x}, \mathbf{x}' \in \mathcal{P}, \mathbf{x} \neq \mathbf{x}'} p_1(\mathbf{x}) p_2(\mathbf{x}')\right] = \int_{\Omega \times \Omega} p_1(\mathbf{x}) p_2(\mathbf{x}') M^{(2)}(d\mathbf{x}, d\mathbf{x}') \quad (35a)$$

$$= \int_{\Omega \times \Omega} p_1(\mathbf{x}) p_2(\mathbf{x}') \phi(\mathbf{x}) \phi(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \quad (35b)$$

$$= \int_{\Omega} p_1(\mathbf{x}) \phi(\mathbf{x}) \underbrace{\left(\int_{\Omega} p_2(\mathbf{x}') \phi(\mathbf{x}') d\mathbf{x}'\right)}_{\text{constant w.r.t. } \mathbf{x}} d\mathbf{x} \quad (35c)$$

$$= \left(\int_{\Omega} p_1(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x}\right) \left(\int_{\Omega} p_2(\mathbf{x}') \phi(\mathbf{x}') d\mathbf{x}'\right) \quad (35d)$$

$$= \mathbb{E}[\mathcal{E}_{p_1}] \mathbb{E}[\mathcal{E}_{p_2}]. \quad (35e)$$

Substituting Eq. (35e) in Eq. (33) yields

$$\mathbb{E}[\mathcal{E}_{p_1} \mathcal{E}_{p_2}] = \langle p_1 p_2, \phi \rangle + \mathbb{E}[\mathcal{E}_{p_1}] \mathbb{E}[\mathcal{E}_{p_2}] \quad (36)$$

Then substituting Eq. (36) in Eq. (32) results in our desired result Eq. (29):

$$\text{Cov}(\mathcal{E}_{p_1}, \mathcal{E}_{p_2}) = \langle p_1 p_2, \phi \rangle + \mathbb{E}[\mathcal{E}_{p_1}] \mathbb{E}[\mathcal{E}_{p_2}] - \mathbb{E}[\mathcal{E}_{p_1}] \mathbb{E}[\mathcal{E}_{p_2}] = \langle p_1 p_2, \phi \rangle. \quad (37)$$

□

## B.4 Noise model for Fourier probing measurements (Equation 2)

In this section, we derive the noise model for our probing measurements distribution described in Equations 2, 3, and 4.

**Equation 2.** Fix spatiotemporal frequency  $\mathbf{f} = [f_x, f_y, f_t]$ . The Fourier basis probing functions take the form  $p_{\mathbf{f}}(\mathbf{x}) = e^{-j2\pi \mathbf{f}^T \mathbf{x}}$ , where the probing measurement  $\mathcal{E}_{\mathbf{f}}$  is defined as

$$\mathcal{E}_{\mathbf{f}} \triangleq \sum_{\mathbf{x} \in \mathcal{P}} e^{-j2\pi \mathbf{f}^T \mathbf{x}}. \quad (2)$$

### B.4.1 Mean and covariance matrix of Fourier probing measurements (proof of Equation 3)

**Equation 3.** For 3D Fourier basis functions, the probing measurements  $\mathcal{E}_f$  approximately follow a complex normal distribution with mean and covariance

$$\boldsymbol{\mu} = \begin{bmatrix} \langle \cos(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle \\ \langle -\sin(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \langle \cos^2(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle & 0 \\ 0 & \langle \sin^2(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle \end{bmatrix}, \quad (3)$$

where the off-diagonal terms are negligible for natural videos by the power law decay of their power spectra.

**Corollary 1.** Fix frequency  $\mathbf{f}$ . Suppose  $2\mathbf{f}$  contributes little signal to  $\phi$  (i.e.,  $\langle p_{2f}, \phi \rangle \approx 0$ ), or constitutes a high frequency in a natural video. Its covariance matrix can be approximated as

$$\boldsymbol{\Sigma} \approx \frac{\langle 1, \phi \rangle}{2} I_2 = \underbrace{\frac{\mathbb{E}[|\mathcal{P}|]}{2}}_{\approx |\mathcal{P}|/2} I_2, \quad (38)$$

where  $I_2$  is the  $2 \times 2$  identity matrix. Equality holds if  $2\mathbf{f}$  contributes no signal to  $\phi$  (i.e.,  $\langle p_{2f}, \phi \rangle = \mathbb{E}[\mathcal{E}_{2f}] = 0$ ).

*Proof Sketch of Equation 3. Step 1:* We separate the real and imaginary parts using Euler's formula, and derive their mean.

**Step 2:** We derive the covariance matrix between the real and imaginary parts of the Fourier probing measurements, and show that for natural videos, it is approximately diagonal, since their power spectra generally follow the power law [8]. This lets us conclude the real and imaginary components are approximately independent normal distributions.  $\square$

*Proof of Equation 3. Step 1:* Euler's formula provides

$$p_f(\mathbf{x}) = \cos(2\pi \mathbf{f}^\top \mathbf{x}) - j \sin(2\pi \mathbf{f}^\top \mathbf{x}). \quad (39)$$

Since  $\text{Re}[p_f(\mathbf{x})] = \cos(2\pi \mathbf{f}^\top \mathbf{x})$  and  $\text{Im}[p_f(\mathbf{x})] = -\sin(2\pi \mathbf{f}^\top \mathbf{x})$ , Proposition 2 gives us

$$\mathbb{E}[\text{Re}[\mathcal{E}_f]] = \langle \cos(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle, \quad (40a)$$

$$\mathbb{E}[\text{Im}[\mathcal{E}_f]] = \langle -\sin(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle. \quad (40b)$$

**Step 2:** Similarly, the variances are given by

$$\text{Var}(\text{Re}[\mathcal{E}_f]) = \langle \cos^2(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle \quad (41a)$$

$$= \frac{1}{2} \underbrace{\langle 1, \phi(\mathbf{x}) \rangle}_{\approx |\mathcal{P}|} + \frac{1}{2} \langle \cos(4\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle, \quad (41b)$$

$$\text{Var}(\text{Im}[\mathcal{E}_f]) = \langle \sin^2(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle \quad (42a)$$

$$= \frac{1}{2} \langle 1, \phi(\mathbf{x}) \rangle - \frac{1}{2} \langle \cos(4\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle. \quad (42b)$$

From Proposition 3, the covariance  $\text{Cov}(\text{Re}[\mathcal{E}_f], \text{Im}[\mathcal{E}_f])$  is given by

$$\text{Cov}(\text{Re}[\mathcal{E}_f], \text{Im}[\mathcal{E}_f]) = -\langle \cos(2\pi \mathbf{f}^\top \mathbf{x}) \sin(2\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle \quad (43a)$$

$$= -\frac{1}{2} \langle \sin(4\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle. \quad (43b)$$

We note that Eq. (43b) is the exact expression for the covariance. Since both  $\langle \sin(4\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle$  and  $\langle \cos(4\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle$  are bounded above by  $A_{2f}/2$ , where  $A_{2f}$  is the amplitude of frequency  $2\mathbf{f}$  in  $\phi$ , if the frequency  $2\mathbf{f}$  is not present in the flux, then  $A_{2f} = 0$  (or equivalently, whenever  $\langle \sin(4\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle = 0$  and  $\langle \cos(4\pi \mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle = 0$ ), and the covariance becomes diagonal.

More generally, when the projection of  $\phi$  onto frequency  $2\mathbf{f}$  is small, these terms are small relative to  $\langle 1, \phi \rangle$ , and the covariance matrix is well-approximated as diagonal. In practice, high-frequency components of  $\phi$  often decay, so the  $2\mathbf{f}$  projections become small as  $\|\mathbf{f}\|$  increases, since natural videos tend to follow the power law [8].

More specifically, in the domain of high temporal frequencies  $f_t$  and low spatial frequencies  $\mathbf{f}_s = (f_x, f_y)$ , the power spectrum  $R(\mathbf{f}_s, f_t)$  follows the asymptotic relation [8]

$$R(\mathbf{f}_s, f_t) \sim \frac{1}{\|\mathbf{f}_s\|^{m-1} f_t^2} \quad (44)$$

In practice, high frequency components  $\mathbf{f}$  often decay quickly, thus their covariances quickly become small. We can approximate  $\langle \sin(4\pi\mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle \approx 0$  and  $\langle \cos(4\pi\mathbf{f}^\top \mathbf{x}), \phi(\mathbf{x}) \rangle \approx 0$ . This means the covariance terms of  $\mathcal{E}_{\mathbf{f}}$  can be approximated as follows:

$$\text{Cov}(\text{Re}[\mathcal{E}_{\mathbf{f}}], \text{Im}[\mathcal{E}_{\mathbf{f}}]) \approx 0, \quad \text{Var}(\text{Re}[\mathcal{E}_{\mathbf{f}}]) \approx \text{Var}(\text{Im}[\mathcal{E}_{\mathbf{f}}]) \approx \frac{\langle 1, \phi \rangle}{2}. \quad (45)$$

Where equality holds if  $2\mathbf{f}$  contributes no energy to  $\phi$ .

From Proposition 2, this distribution is approximately normal.  $\square$

## B.4.2 Fourier probing energy (proof of Equation 4)

**Equation 4.** The total normalized Fourier probing energy is given by

$$|\tilde{\mathcal{E}}_{\mathbf{f}}|^2 \triangleq \text{Re} \left( \frac{\mathcal{E}_{\mathbf{f}}}{\sqrt{\Sigma_{1,1}}} \right)^2 + \text{Im} \left( \frac{\mathcal{E}_{\mathbf{f}}}{\sqrt{\Sigma_{2,2}}} \right)^2. \quad (4)$$

Under the diagonal covariance described in Equation 3, it follows a noncentral  $\chi^2$  distribution with two degrees of freedom.

**Corollary 2** (Probing energy under the null hypothesis). Under the null hypothesis that both  $\mathbf{f}$  and  $2\mathbf{f}$  contribute no energy to  $\phi$  (i.e.,  $\langle p_{\mathbf{f}}, \phi \rangle = 0$  and  $\langle p_{2\mathbf{f}}, \phi \rangle = 0$ ), the covariance matrix reduces to  $\Sigma = \frac{\langle 1, \phi \rangle}{2} I_2$ , which means

$$|\tilde{\mathcal{E}}_{\mathbf{f}}|^2 = \frac{2}{\langle 1, \phi \rangle} |\mathcal{E}_{\mathbf{f}}|^2. \quad (46)$$

Thus,  $|\mathcal{E}_{\mathbf{f}}|^2$  follows a scaled  $\chi^2$  distribution with two degrees of freedom:

$$|\mathcal{E}_{\mathbf{f}}|^2 \sim \frac{\langle 1, \phi \rangle}{2} \chi_2^2 \approx \frac{|\mathcal{P}|}{2} \chi_2^2 \quad (47)$$

*Proof of Equation 4.*  $|\tilde{\mathcal{E}}_{\mathbf{f}}|^2$  is the sum of two independent squared normal distributions with unit variance, which by definition follows a noncentral  $\chi^2$  distribution with noncentrality parameter  $\lambda$ :

$$|\tilde{\mathcal{E}}_{\mathbf{f}}|^2 \sim \chi_2^2(\lambda), \quad \lambda = \frac{\mu_1^2}{\Sigma_{1,1}} + \frac{\mu_2^2}{\Sigma_{2,2}}. \quad (48)$$

Under the null hypothesis,  $\lambda = 0$  and the distribution is central  $\chi_2^2$ .  $\square$

## B.5 CFAR frequency detection (proof of Equation 5)

**Equation 5.** We use normalized probing function  $p_{\mathbf{f}}(\mathbf{x}) = \frac{1}{\sqrt{v}} e^{-j2\pi\mathbf{f}^\top \mathbf{x}}$ , where  $v = |\Omega| = w \cdot h \cdot t_{\text{exp}}$  is the spatiotemporal volume of  $\Omega$ , such that they are orthonormal over  $\Omega$ . Then frequency is selected with significance level  $\alpha$  if

$$|\mathcal{E}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \frac{|\mathcal{P}|}{2v}. \quad (5)$$

*Proof of Equation 5.* Under the null hypothesis that  $\mathbf{f}$  and  $2\mathbf{f}$  both contribute no energy to  $\phi$ , the energy  $|\mathcal{E}_{\mathbf{f}}|^2$  follows a scaled central  $\chi_2^2$  distribution from Eq. (47), where the scale factor includes the spatiotemporal volume  $v$  from the normalized

probing function:

$$|\mathcal{E}_{\mathbf{f}}|^2 \sim \frac{\langle 1, \phi \rangle}{2v} \chi_2^2. \quad (49)$$

When  $\langle p_{2\mathbf{f}}, \phi \rangle \neq 0$ , the covariance is no longer isotropic and the test is approximate. However, since natural video power spectra decay as a power law with frequency, energy at  $2\mathbf{f}$  is suppressed relative to  $\mathbf{f}$ , and the isotropic approximation holds well in practice (see Section I.1).

The exact detection threshold  $\text{CDF}_{\chi_2^2}^{-1}(1 - \alpha)^{\frac{\langle 1, \phi \rangle}{2v}}$  depends on the unknown flux  $\phi$  and is therefore unobservable. Approximating via the observed photon count  $|\mathcal{P}| \approx \mathbb{E}[|\mathcal{P}|] = \langle 1, \phi \rangle$  gives Equation 5. We show in Section I.2 that this approximation performs well in practice.  $\square$

## C Velocity Detection

In this section, we provide the derivations and proofs for the velocity detection framework presented in Section 3.2 of the main paper.

### C.1 Fourier planar structures induced by linear velocity (proof of Equation 6)

For convenience, we restate Equation 6 from the main paper.

**Equation 6.** Let a local patch of spatial intensity  $I$  translate at a constant velocity  $\mathbf{v} = [v_x, v_y]^\top$  such that

$$\phi(x, y, t) = I(x - v_x t, y - v_y t).$$

Then the Fourier coefficients lie on the plane

$$v_x f_x + v_y f_y + f_t = 0 \quad (6)$$

*Proof Sketch of Equation 6.* The proof proceeds in two steps.

**Step 1.** We write the Fourier transform of the translated pattern as a space-time integral and perform a change of variables that aligns the coordinates with the motion.

**Step 2.** We evaluate the transformed integral and show that it vanishes unless the frequencies satisfy Equation 6

□

*Proof of Equation 6.*

**Step 1.** Let

$$\phi(x, y, t) = I(x - v_x t, y - v_y t), \quad (50)$$

for  $(x, y, t) \in \mathbb{R}^3$ .

The Fourier transform of  $\phi(x, y, t)$  is

$$\Phi(f_x, f_y, f_t) = \iiint_{(x,y,t) \in \mathbb{R}^3} \phi(x, y, t) e^{-i2\pi(f_x x + f_y y + f_t t)} dx dy dt. \quad (51)$$

Substituting Eq. 50 into Eq. 51 gives

$$\Phi(f_x, f_y, f_t) = \iiint_{(x,y,t) \in \mathbb{R}^3} I(x - v_x t, y - v_y t) e^{-i2\pi(f_x x + f_y y + f_t t)} dx dy dt. \quad (52)$$

We now introduce the change of variables

$$\xi = x - v_x t, \quad \eta = y - v_y t, \quad \tau = t, \quad (53)$$

for  $(\xi, \eta, \tau) \in \mathbb{R}^3$ . Equivalently,

$$x = \xi + v_x \tau, \quad y = \eta + v_y \tau, \quad t = \tau. \quad (54)$$

The Jacobian matrix of this transformation is

$$\frac{\partial(x, y, t)}{\partial(\xi, \eta, \tau)} = \begin{pmatrix} 1 & 0 & v_x \\ 0 & 1 & v_y \\ 0 & 0 & 1 \end{pmatrix}, \quad (55)$$

whose determinant is

$$\left| \frac{\partial(x, y, t)}{\partial(\xi, \eta, \tau)} \right| = 1. \quad (56)$$

Therefore,

$$dx dy dt = d\xi d\eta d\tau. \quad (57)$$

Applying the change of variables Eq. 53–Eq. 57 to Eq. 52, we obtain

$$\Phi(f_x, f_y, f_t) = \iiint_{(\xi, \eta, \tau) \in \mathbb{R}^3} I(\xi, \eta) e^{-i2\pi[f_x(\xi+v_x\tau)+f_y(\eta+v_y\tau)+f_t\tau]} d\xi d\eta d\tau. \quad (58)$$

Rearranging the exponent gives

$$\Phi(f_x, f_y, f_t) = \iiint_{(\xi, \eta, \tau) \in \mathbb{R}^3} I(\xi, \eta) e^{-i2\pi(f_x\xi+f_y\eta)} e^{-i2\pi(v_x f_x + v_y f_y + f_t)\tau} d\xi d\eta d\tau. \quad (59)$$

**Step 2.** The integrand in Eq. 59 separates into a spatial term depending only on  $(\xi, \eta)$  and a temporal term depending only on  $\tau$ . Hence,

$$\Phi(f_x, f_y, f_t) = \left( \iint_{(\xi, \eta) \in \mathbb{R}^2} I(\xi, \eta) e^{-i2\pi(f_x\xi+f_y\eta)} d\xi d\eta \right) \left( \int_{\tau \in \mathbb{R}} e^{-i2\pi(v_x f_x + v_y f_y + f_t)\tau} d\tau \right). \quad (60)$$

Define

$$\widehat{I}(f_x, f_y) = \iint_{(\xi, \eta) \in \mathbb{R}^2} I(\xi, \eta) e^{-i2\pi(f_x\xi+f_y\eta)} d\xi d\eta \quad (61)$$

and

$$J(f_x, f_y, f_t) = \int_{\tau \in \mathbb{R}} e^{-i2\pi(v_x f_x + v_y f_y + f_t)\tau} d\tau. \quad (62)$$

Then Eq. 60 becomes

$$\Phi(f_x, f_y, f_t) = \widehat{I}(f_x, f_y) J(f_x, f_y, f_t). \quad (63)$$

Using the distributional identity

$$\int_{\tau \in \mathbb{R}} e^{-i2\pi\alpha\tau} d\tau = \delta(\alpha), \quad (64)$$

with

$$\alpha = v_x f_x + v_y f_y + f_t, \quad (65)$$

we obtain

$$J(f_x, f_y, f_t) = \delta(v_x f_x + v_y f_y + f_t). \quad (66)$$

Substituting Eq. 66 into Eq. 63 yields

$$\Phi(f_x, f_y, f_t) = \widehat{I}(f_x, f_y) \delta(v_x f_x + v_y f_y + f_t). \quad (67)$$

Therefore,  $\Phi(f_x, f_y, f_t)$  can be nonzero only when

$$v_x f_x + v_y f_y + f_t = 0. \quad (68)$$

That is, the 3D Fourier coefficients of  $\phi$  are supported on a plane through the origin with normal vector  $[v_x, v_y, 1]^\top$ . This concludes the proof.  $\square$

**Remark** (Finite observation window). The identity in Eq. 64 assumes integration over  $\tau \in \mathbb{R}$ . For a finite observation interval  $\tau \in [0, t_{\text{exp}}]$ , the delta distribution is replaced by a sinc-like kernel of width on the order of  $1/t_{\text{exp}}$ , and the Fourier energy concentrates near the plane Eq. 68 rather than exactly on it. Thus, the plane constraint is exact in the limit  $t_{\text{exp}} \rightarrow \infty$  and approximate for finite  $t_{\text{exp}}$ .

## C.2 Parametric velocity detection model

We next describe a parametric CFAR model for the local velocity energies. Let  $\Pi_i$  denote the set of Fourier coefficients lying on the candidate plane associated with velocity hypothesis  $\mathbf{v}_i$ , and let

$$E_i \triangleq \sum_{\mathbf{f} \in \Pi_i} |\mathcal{E}_{\mathbf{f}}|^2 \quad (69)$$

be the corresponding plane energy. Under the null hypothesis  $H_0$ , assume that the estimate of the Fourier coefficients on the plane are independent circular complex normal random variables with zero mean,

$$\mathcal{E}_{\mathbf{f}} \mid H_0 \sim \mathcal{CN}(0, \sigma_i^2), \quad \mathbf{f} \in \Pi_i. \quad (70)$$

Writing each estimated coefficient as

$$\mathcal{E}_{\mathbf{f}} = X_{\mathbf{f}} + \mathbf{i}Y_{\mathbf{f}}, \quad (71)$$

with

$$X_{\mathbf{f}}, Y_{\mathbf{f}} \sim \mathcal{N}(0, \sigma_i^2/2), \quad (72)$$

we have

$$|\mathcal{E}_{\mathbf{f}}|^2 = X_{\mathbf{f}}^2 + Y_{\mathbf{f}}^2, \quad (73)$$

and therefore

$$\frac{2|\mathcal{E}_{\mathbf{f}}|^2}{\sigma_i^2} \sim \chi_2^2. \quad (74)$$

If the plane contains  $|\Pi_i|$  estimated coefficients, then summing over all the ones on the plane gives

$$\frac{2E_i}{\sigma_i^2} = \sum_{\mathbf{f} \in \Pi_i} \frac{2|\mathcal{E}_{\mathbf{f}}|^2}{\sigma_i^2} \sim \chi_{2|\Pi_i|}^2. \quad (75)$$

Hence, under  $H_0$ , the plane energy  $E_i$  follows a scaled central chi-squared distribution with  $2|\Pi_i|$  degrees of freedom.

Under the alternative hypothesis  $H_1$ , the plane contains true signal energy, so the estimated Fourier coefficients have nonzero means:

$$\mathcal{E}_{\mathbf{f}} \sim \mathcal{CN}(\mu_{\mathbf{f}}, \sigma_i^2). \quad (76)$$

In this case,

$$\frac{2E_i}{\sigma_i^2} \sim \chi_{2|\Pi_i|}^2(\lambda_i), \quad \lambda_i = \sum_{\mathbf{f} \in \Pi_i} \frac{2|\mu_{\mathbf{f}}|^2}{\sigma_i^2}, \quad (77)$$

so that  $E_i$  follows a scaled noncentral chi-squared distribution [21].

**Remark** (Gamma and Gaussian forms under  $H_0$ ). Since

$$\frac{1}{2}\chi_{2k}^2 \sim \text{Gamma}(k, 1), \quad (78)$$

the null distribution can equivalently be written as

$$E_i \sim \text{Gamma}(|\Pi_i|, \sigma_i^2), \quad (79)$$

with

$$\mathbb{E}[E_i \mid H_0] = |\Pi_i|\sigma_i^2, \quad \mathbf{Var}[E_i \mid H_0] = |\Pi_i|\sigma_i^4. \quad (80)$$

As  $|\Pi_i|$  grows, the classical CLT for i.i.d. summands gives the normal approximation

$$E_i \mid H_0 \approx \mathcal{N}(|\Pi_i|\sigma_i^2, |\Pi_i|\sigma_i^4). \quad (81)$$

Therefore, when the number of coefficients accumulated on the plane is large, the null distribution of the plane energy may be approximated by a normal distribution with the mean and variance above.

### C.3 Limitation of the parametric velocity detection model

We next discuss the assumptions underlying a parametric velocity-plane detector and motivate the use of a nonparametric alternative. Implementing a parametric CFAR detector would require estimating at least the local scale parameter  $\sigma_i^2$ , and in the nonzero-mean case also the noncentrality parameter  $\lambda_i$ , from neighboring training cells.

A natural approach under this model is to use a window-based CFAR detector, such as CA-CFAR or GO-CFAR, in which a local neighborhood is defined around each candidate velocity  $\mathbf{v}_i$ , a guard region is excluded around the cell under test, and the remaining cells are used to estimate the local background level and set a detection threshold. This yields a parametric test whose false alarm probability is controlled only insofar as the assumed distributional model is accurate.

However, these assumptions are often too strong in our setting. In particular, the Fourier coefficients accumulated along nearby velocity planes are generally not independent, since nearby planes overlap substantially in Fourier space and are computed from the same underlying data. As a result, nearby plane energies are also correlated, and the assumption that the local plane energies are independent and identically distributed can be difficult to justify. For this reason, a fully parametric CFAR model is brittle when applied to the velocity detection problem.

This motivates a weaker nonparametric null model. Rather than assuming that the local plane energies are independent and identically distributed, we assume only that under  $H_0$  the collection

$$E_i, \{E_j\}_{j \in \mathcal{N}_i}$$

is exchangeable and has a continuous joint distribution [30]. Exchangeability means that the joint distribution is invariant under any permutation of the variables, so that under the null hypothesis no candidate in the local neighborhood is statistically privileged over any other. Continuity ensures that ties occur with probability zero. These assumptions are sufficient to derive the uniform rank law used by the rank-based CFAR detector, while avoiding the stronger and less realistic requirement of independence among neighboring plane energies.

### C.4 Nonparametric velocity detection model (proof of Equation 8)

In this section, we prove Equation 8 of the main paper, after providing notation.

Let  $\mathbf{v}_i = [v_x, v_y]^\top$  denote a candidate velocity hypothesis in a discrete velocity grid, and let  $E_i$  be its associated energy. For each hypothesis  $\mathbf{v}_i$ , we define a local neighborhood  $\mathcal{N}_i$  by centering a square window around  $\mathbf{v}_i$  and excluding a smaller guard region around the cell under test. The guard region prevents the energy of the tested hypothesis and its immediate neighbors from contaminating the background set, while the remaining cells in the window serve as neighboring hypotheses for comparison. We denote the number of such neighboring hypotheses by  $|\mathcal{N}_i|$ .

**Equation 8.** Let  $E_i$  denote the energy of a candidate velocity hypothesis  $\mathbf{v}_i$ , and let  $\mathcal{N}_i$  be its set of neighboring hypotheses used as training cells. Define the rank

$$r_i \triangleq |\{j \in \mathcal{N}_i : E_j < E_i\}|.$$

Then detecting  $E_i$  by requiring it to lie in the top  $\alpha_{\text{vel}}$  fraction of its local neighborhood is equivalent to the test

$$r_i \geq \lceil (1 - \alpha_{\text{vel}})(|\mathcal{N}_i| + 1) \rceil. \quad (8)$$

Moreover, under the null hypothesis  $H_0$  that the collection

$$E_i, \{E_j\}_{j \in \mathcal{N}_i}$$

is exchangeable and has a continuous joint distribution, the rank  $r_i$  is uniformly distributed on  $\{0, 1, \dots, |\mathcal{N}_i|\}$ . Consequently, the false alarm probability of the detector depends only on the selected threshold and the neighborhood size, thereby yielding a CFAR test.

*Proof sketch of Equation 8.* The proof proceeds in four steps:

**Step 1:** We relate the rank  $r_i$  to the position of  $E_i$  in the sorted list formed by the cell under test together with its neighboring hypotheses.

**Step 2:** We show that requiring  $E_i$  to belong to the top  $\alpha_{\text{vel}}$  fraction of that list is equivalent to the threshold in Equation 8.

**Step 3:** Under  $H_0$ , we show that exchangeability implies that all positions of the cell under test in the sorted list are equally likely, while continuity guarantees that ties occur with probability zero.

**Step 4:** We compute the probability of detection under  $H_0$  and show that it depends only on the threshold and the neighborhood size.  $\square$

*Proof of Equation 8.*

**Step 1:** Consider the set consisting of the cell under test and its neighboring hypotheses,

$$\{E_i\} \cup \{E_j : j \in \mathcal{N}_i\}. \quad (82)$$

Its total size is

$$|\mathcal{N}_i| + 1. \quad (83)$$

Let these values be sorted in ascending order. Since

$$r_i = |\{j \in \mathcal{N}_i : E_j < E_i\}|, \quad (84)$$

the quantity  $r_i$  counts exactly how many neighboring energies are strictly smaller than  $E_i$ . Therefore, the position of  $E_i$  in the ascending ordering is

$$r_i + 1. \quad (85)$$

**Step 2:** We require that  $E_i$  lies in the top  $\alpha_{\text{vel}}$  fraction of the  $|\mathcal{N}_i| + 1$  values. This means that the ascending position of  $E_i$  must satisfy

$$r_i + 1 > (1 - \alpha_{\text{vel}})(|\mathcal{N}_i| + 1), \quad (86)$$

*i.e.*,

$$r_i > (1 - \alpha_{\text{vel}})(|\mathcal{N}_i| + 1) - 1. \quad (87)$$

Since  $r_i$  is integer-valued, this is equivalent to

$$r_i \geq \lceil (1 - \alpha_{\text{vel}})(|\mathcal{N}_i| + 1) \rceil. \quad (88)$$

This is exactly Eq. 8. Consequently, detecting a hypothesis only when its energy lies in the top  $\alpha_{\text{vel}}$  fraction of the combined set is equivalent to thresholding its rank by Eq. 8.

**Step 3:** Under the null hypothesis  $H_0$ , the collection

$$E_i, \{E_j\}_{j \in \mathcal{N}_i} \quad (89)$$

is exchangeable and has a continuous joint distribution. Since the joint distribution is continuous, ties occur with probability zero. Therefore, with probability one, these values can be arranged in a strict ascending order.

Because the collection is exchangeable, its joint distribution is invariant under permutation of the samples. Hence, no sample is favored over any other in the ordering, and  $E_i$  is equally likely to occupy any of the  $|\mathcal{N}_i| + 1$  positions in the ascending ordering. For any fixed  $k \in \{0, 1, \dots, |\mathcal{N}_i|\}$ , the event  $r_i = k$  means that exactly  $k$  neighboring energies are smaller than  $E_i$ , or equivalently, that  $E_i$  occupies position  $k + 1$  in the ascending ordering. Since each position is equally likely, we have

$$\Pr(r_i = k \mid H_0) = \frac{1}{|\mathcal{N}_i| + 1}, \quad k = 0, 1, \dots, |\mathcal{N}_i|. \quad (90)$$

Therefore, under  $H_0$ , the rank  $r_i$  is uniformly distributed on  $\{0, 1, \dots, |\mathcal{N}_i|\}$  [25].

**Step 4:** Let

$$T_i = \lceil (1 - \alpha_{\text{vel}})(|\mathcal{N}_i| + 1) \rceil. \quad (91)$$

Under  $H_0$ , the detector declares a detection whenever  $r_i \geq T_i$ . Since  $r_i$  is uniformly distributed on  $\{0, 1, \dots, |\mathcal{N}_i|\}$ , we obtain

$$\Pr(\text{detect} \mid H_0) = \Pr(r_i \geq T_i \mid H_0) = \sum_{k=T_i}^{|\mathcal{N}_i|} \Pr(r_i = k \mid H_0) = \sum_{k=T_i}^{|\mathcal{N}_i|} \frac{1}{|\mathcal{N}_i| + 1}. \quad (92)$$

Evaluating the sum gives

$$\Pr(\text{detect} \mid H_0) = \frac{|\mathcal{N}_i| - T_i + 1}{|\mathcal{N}_i| + 1}. \quad (93)$$

Substituting the definition of  $T_i$ , we obtain

$$\mathbf{Pr}(\text{detect} \mid H_0) = \frac{|\mathcal{N}_i| - \lceil (1 - \alpha_{\text{vel}})(|\mathcal{N}_i| + 1) \rceil + 1}{|\mathcal{N}_i| + 1}. \quad (94)$$

Therefore, under the null hypothesis, the false alarm probability depends only on the selected threshold and the neighborhood size, and not on the underlying background distribution. This establishes the CFAR property and concludes the proof.  $\square$

**Remark** (Exact false alarm level on integer-compatible grids). In general, the false alarm probability is

$$\mathbf{Pr}(\text{detect} \mid H_0) = \frac{|\mathcal{N}_i| - \lceil (1 - \alpha_{\text{vel}})(|\mathcal{N}_i| + 1) \rceil + 1}{|\mathcal{N}_i| + 1}. \quad (95)$$

In particular, this expression reduces exactly to

$$\mathbf{Pr}(\text{detect} \mid H_0) = \alpha_{\text{vel}} \quad (96)$$

whenever  $\alpha_{\text{vel}}(|\mathcal{N}_i| + 1)$  is an integer. Indeed, in that case,

$$\lceil (1 - \alpha_{\text{vel}})(|\mathcal{N}_i| + 1) \rceil = (|\mathcal{N}_i| + 1) - \alpha_{\text{vel}}(|\mathcal{N}_i| + 1), \quad (97)$$

and substituting into Eq. 95 gives

$$\mathbf{Pr}(\text{detect} \mid H_0) = \frac{|\mathcal{N}_i| - ((|\mathcal{N}_i| + 1) - \alpha_{\text{vel}}(|\mathcal{N}_i| + 1)) + 1}{|\mathcal{N}_i| + 1} = \alpha_{\text{vel}}. \quad (98)$$

Therefore, when  $\alpha_{\text{vel}}(|\mathcal{N}_i| + 1)$  is integer-valued, the rank threshold yields exactly the desired false alarm probability.

## D Extension to Other Asynchronous Sensors

### D.1 Spike Cameras: Spatiotemporal Probing Derivation

In this section, we derive the extension of spatiotemporal flux probing to spike cameras. We start with some notation and derive the extension to 1D temporal probing.

**Integrate-and-fire mechanism:** The spike camera pixel integrates the photon arrivals into an internal accumulator. A spike is emitted each time the accumulator reaches a threshold  $\Theta > 0$  (in units of photon counts), and the accumulator resets to zero [16]. Let  $\{\tau_k(\mathbf{r})\}_{k=1}^{N^{\text{sp}}(\mathbf{r})}$  be the spike times, with  $\tau_0(\mathbf{r}) \triangleq 0$ . Define the spike counting process

$$N_{\mathbf{r}}^{\text{sp}}(t) \triangleq |\{k : \tau_k(\mathbf{r}) \leq t\}|. \quad (99)$$

By construction, the  $k$ -th spike fires when exactly  $\Theta$  photons have arrived since the  $(k-1)$ -th spike:

$$N_{\mathbf{r}}^{\text{ph}}(\tau_k(\mathbf{r})) - N_{\mathbf{r}}^{\text{ph}}(\tau_{k-1}(\mathbf{r})) = \Theta, \quad k = 1, 2, \dots \quad (100)$$

and the residual photon count since the last spike satisfies

$$\varepsilon_{\mathbf{r}}(t) \triangleq N_{\mathbf{r}}^{\text{ph}}(t) - N_{\mathbf{r}}^{\text{ph}}(\tau_{N_{\mathbf{r}}^{\text{sp}}(t)}(\mathbf{r})) < \Theta. \quad (101)$$

This gives the fundamental identity

$$N_{\mathbf{r}}^{\text{ph}}(t) = \Theta N_{\mathbf{r}}^{\text{sp}}(t) + \varepsilon_{\mathbf{r}}(t). \quad (102)$$

#### D.1.1 Temporal spike probing

**Proposition 4** (Spike probing decomposition). Let  $p$  be a deterministic, bounded, and differentiable probing function. Then

$$\Theta \sum_{k=1}^{N^{\text{sp}}(\mathbf{r})} p(\tau_k(\mathbf{r})) = \int_0^{t_{\text{exp}}} p(t) \phi(\mathbf{r}, t) dt + M_{p,\mathbf{r}} + \eta_{p,\mathbf{r}}, \quad (103)$$

where:

- $M_{p,\mathbf{r}}$  is martingale noise inherited from Eq. (20), the flux probing measurements,
- $\eta_{p,\mathbf{r}}$  is a bounded quantization residual satisfying

$$|\eta_{p,\mathbf{r}}| \leq \Theta(|p(t_{\text{exp}})| + |p(0)|) + \Theta \int_0^{t_{\text{exp}}} |p'(t)| dt. \quad (104)$$

*Proof Sketch of Proposition 4. Step 1:* We substitute the 1D martingale decomposition from the photon counting process described in Eq. (12), and show it can be represented in differential form.

**Step 2:** We integrate the differential form of the decomposition from step 1, and evaluate the individual terms.

**Step 3:** We derive the bounds of the quantization error.  $\square$

*Proof. Step 1:* The martingale decomposition described in Eq. (12) can be substituted directly into the photon-spike identity Eq. 102:

$$\int_0^t \phi(\mathbf{r}, s) ds + M_{\mathbf{r}}(t) = \Theta N_{\mathbf{r}}^{\text{sp}}(t) + \varepsilon_{\mathbf{r}}(t). \quad (105)$$

Both sides of Eq. 105 are functions of  $t$  that change only by jumps (at photon arrivals on the left, at spike emissions and photon arrivals on the right) or by smooth accumulation (the flux integral on the left). In particular, every term is a well-defined increasing or piecewise-constant process whose value at any time is determined by finitely many photon and spike events up to that time. This regularity is sufficient to take increments on both sides [4], giving the differential form

$$\phi(\mathbf{r}, t) dt + dM_{\mathbf{r}}(t) = \Theta dN_{\mathbf{r}}^{\text{sp}}(t) + d\varepsilon_{\mathbf{r}}(t). \quad (106)$$

**Step 2:** Multiply both sides of Eq. 106 by  $p(t)$  and integrate over  $[0, t_{\text{exp}}]$ :

$$\int_0^{t_{\text{exp}}} p(t) \phi(\mathbf{r}, t) dt + \int_0^{t_{\text{exp}}} p(t) dM_{\mathbf{r}}(t) = \Theta \int_0^{t_{\text{exp}}} p(t) dN_{\mathbf{r}}^{\text{sp}}(t) + \int_0^{t_{\text{exp}}} p(t) d\varepsilon_{\mathbf{r}}(t). \quad (107)$$

Since  $dN_{\mathbf{r}}^{\text{sp}}(t) = \sum_k \delta(t - \tau_k(\mathbf{r})) dt$ , the integral  $\int_0^{t_{\text{exp}}} p(t) dN_{\mathbf{r}}^{\text{sp}}(t)$  evaluates  $p$  along a step process, which evaluates to the sum

$$\int_0^{t_{\text{exp}}} p(t) dN_{\mathbf{r}}^{\text{sp}}(t) = \sum_{k=1}^{N^{\text{sp}}(\mathbf{r})} p(\tau_k(\mathbf{r})). \quad (108)$$

Rearranging Eq. 107 gives

$$\Theta \sum_{k=1}^{N^{\text{sp}}(\mathbf{r})} p(\tau_k(\mathbf{r})) = \int_0^{t_{\text{exp}}} p(t) \phi(\mathbf{r}, t) dt + \underbrace{\int_0^{t_{\text{exp}}} p(t) dM_{\mathbf{r}}(t)}_{M_{p,\mathbf{r}}} + \underbrace{\left( - \int_0^{t_{\text{exp}}} p(t) d\varepsilon_{\mathbf{r}}(t) \right)}_{\eta_{p,\mathbf{r}}}, \quad (109)$$

such that  $M_{p,\mathbf{r}}$  is a martingale by Lemma 2.

**Step 3:** Integration by parts gives

$$\begin{aligned} \eta_{p,\mathbf{r}} &= - \int_0^{t_{\text{exp}}} p(t) d\varepsilon_{\mathbf{r}}(t) \\ &= - \left[ p(t_{\text{exp}}) \varepsilon_{\mathbf{r}}(t_{\text{exp}}) - p(0) \varepsilon_{\mathbf{r}}(0) \right] + \int_0^{t_{\text{exp}}} \varepsilon_{\mathbf{r}}(t) p'(t) dt. \end{aligned} \quad (110)$$

Taking absolute values and using  $|\varepsilon_{\mathbf{r}}(t)| < \Theta$ :

$$|\eta_{p,\mathbf{r}}| \leq \Theta (|p(t_{\text{exp}})| + |p(0)|) + \Theta \int_0^{t_{\text{exp}}} |p'(t)| dt. \quad (111) \quad \square$$

**Remark** (Limiting regimes). The decomposition Eq. 103 unifies the photon and spike probing frameworks:

1. **Single-photon limit** ( $\Theta = 1$ ): Every photon triggers a spike, so  $N^{\text{sp}} = N^{\text{ph}}$ , the residual  $\varepsilon_{\mathbf{r}}(t) \equiv 0$  (since it can only be 0), and  $\eta_{p,\mathbf{r}} = 0$ . The equation reduces to the photon flux probing equation:

$$\sum_k p(\tau_k) = \int_0^{t_{\text{exp}}} p(t) \phi(\mathbf{r}, t) dt + M_{p,\mathbf{r}}. \quad (112)$$

2. **Deterministic (high-flux) limit:** When the mean photon count per threshold window  $\Theta/\bar{\phi}$  is large, each inter-spike interval concentrates around its deterministic value by the law of large numbers. The spike times become approximately deterministic functionals of  $\phi$ , and  $M_{p,\mathbf{r}}/\Theta \rightarrow 0$  in a suitable sense. The equation reduces to the following spike probing equation:

$$\Theta \sum_k p(\tau_k) = \int_0^{t_{\text{exp}}} p(t) \phi(\mathbf{r}, t) dt + \eta_{p,\mathbf{r}}. \quad (113)$$

## D.1.2 Spatiotemporal spike probing

In this section, we extend the temporal analysis to the spatiotemporal domain.

**Proposition 5** (Spatiotemporal unified spike probing). Let  $\mathcal{S} = \{(\mathbf{r}_i, t_i)\}_{i=1}^N$  be the full spike event stream over the sensor domain  $A$  and time interval  $[0, t_{\text{exp}}]$ , with each pixel following an independent integrate-and-fire model driven by its own Poisson photon stream with common threshold  $\Theta$ . For any probing function  $p(\mathbf{r}, t)$  that is deterministic, bounded, and

differentiable in  $t$  for each  $\mathbf{r}$ :

$$\Theta \sum_{i=1}^N p(\mathbf{r}_i, t_i) = \int_A \int_0^{t_{\text{exp}}} p(\mathbf{r}, t) \phi(\mathbf{r}, t) dt d\mathbf{r} + M_p + \eta_p, \quad (114)$$

where  $|\eta_p| \leq \int_A \left[ \Theta(|p(\mathbf{r}, t_{\text{exp}})| + |p(\mathbf{r}, 0)|) + \Theta \int_0^{t_{\text{exp}}} |\partial_t p(\mathbf{r}, t)| dt \right] d\mathbf{r}$  and  $M_p$  is the martingale inherited from Equation 1.

*Proof.* Apply Proposition 4 at each pixel  $\mathbf{r}$  using the temporal probe  $t \mapsto p(\mathbf{r}, t)$ , then integrate over  $\mathbf{r} \in A$ . From Lemma 2,  $M_p$  is a martingale.  $\square$

### D.1.3 Fourier probing: null distribution

We now restrict to Fourier probing functions

$$p_{\mathbf{f}}(\mathbf{r}, t) = e^{-j2\pi \mathbf{f}^\top [\mathbf{r}, t]}, \quad (115)$$

where  $\mathbf{f} = (f_x, f_y, f_t)$  is the spatiotemporal frequency.

From Eq. 114, the (scaled) spike Fourier probing measurement is

$$\mathcal{E}_{\mathbf{f}} = \Theta \sum_{i=1}^N e^{-j2\pi \mathbf{f}^\top [\mathbf{r}_i, t_i]} = \underbrace{\langle p_{\mathbf{f}}, \phi \rangle}_{\text{flux probing integral}} + M_{p_{\mathbf{f}}} + \eta_{p_{\mathbf{f}}}, \quad (116)$$

To make the CFAR analysis tractable, we work in the regime where the quantization residual  $\eta_{p_{\mathbf{f}}}$  is negligible relative to the stochastic term. This is justified when:

- The threshold  $\Theta$  is moderate (not too large), so the quantization error is small, or
- A low-pass pre-filter has been applied to suppress the  $O(\Theta|f_t|)$  growth of  $\eta_{p_{\mathbf{f}}}$  at high frequencies (as noted in the remark on periodic spike trains).

**Proposition 6.** Under the null hypothesis  $H_0$ , the flux  $\phi$  has no energy at frequency  $\mathbf{f}$  and  $2\mathbf{f}$ , thus  $\mathcal{E}_{\mathbf{f}}$  approximates a complex normal distribution with

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \frac{\bar{N}}{2} I_2. \quad (117)$$

*Proof.* Under this negligible quantization error  $\eta_{p_{\mathbf{f}}}$ ,  $\mathcal{E}_{\mathbf{f}}$  approximately follows the probing measurements derived in Equation 1, thus the distribution of  $\mathcal{E}_{\mathbf{f}}$  can be approximated by Equation 3.

Since  $\mathbf{f}$  does not contribute to  $\phi$ , the flux probing integral  $\langle p_{\mathbf{f}}, \phi \rangle = 0$ . Evaluating Equation 3 using this integral yields  $\mathbb{E}[\mathcal{E}_{\mathbf{f}}] = 0$ . Corollary 1 provides our covariance  $\boldsymbol{\Sigma} = (\bar{N}/2)I_2$  under those same conditions, where  $\bar{N} = \langle 1, \phi \rangle$  is the expected number of photon arrivals.  $\square$

### D.1.4 CFAR detector for spike cameras

Here, we derive the CFAR detector for spike cameras.

**Corollary 3** (Normalized probing energy under  $H_0$ ). Define the normalized spike probing energy as follows from the normalized Fourier flux probing energy in Equation 4:

$$|\tilde{\mathcal{E}}_{\mathbf{f}}|^2 = \text{Re} \left( \frac{\mathcal{E}_{\mathbf{f}}}{\sqrt{\boldsymbol{\Sigma}_{1,1}}} \right)^2 + \text{Im} \left( \frac{\mathcal{E}_{\mathbf{f}}}{\sqrt{\boldsymbol{\Sigma}_{2,2}}} \right)^2, \quad (118)$$

Since we assume  $H_0$ , then  $|\tilde{\mathcal{E}}_{\mathbf{f}}|^2 \sim \chi_2^2$ . Substituting in the covariance and mean from Proposition 6

$$|\tilde{\mathcal{E}}_{\mathbf{f}}|^2 = \frac{2}{\bar{N}} |\mathcal{E}_{\mathbf{f}}|^2, \quad (119)$$

**Proposition 7** (CFAR detector for spike camera Fourier probing). Given a false alarm probability  $\alpha$ , we detect spatiotemporal frequency  $\mathbf{f} \neq 0$  (reject  $H_0$ ) when

$$|\mathcal{E}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \cdot \frac{\bar{N}}{2}. \quad (120)$$

Since  $\bar{N}$  (the expected total photon count) is not directly observed, the conservative CFAR bound is used instead:

$$|\mathcal{E}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \cdot \frac{\Theta N + \Theta |A|}{2}, \quad (121)$$

where  $N$  is the observed number of spikes.

*Proof of Proposition 7.* The CFAR detection threshold for  $|\tilde{\mathcal{E}}_{\mathbf{f}}|^2$  is

$$|\tilde{\mathcal{E}}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha). \quad (122)$$

Substituting Eq. (119) yields Eq. (120).

To determine a conservative bound, we have the identity Eq. 102 evaluated at  $t = t_{\text{exp}}$ :

$$N_{\text{total}}^{\text{ph}} = \Theta \cdot N_{\text{total}}^{\text{sp}} + \varepsilon_{\text{residual}}, \quad (123)$$

where  $N_{\text{total}}^{\text{sp}} = \sum_{\mathbf{r}} N_{\mathbf{r}}^{\text{sp}}(t_{\text{exp}}) = N$  is the total spike count and  $\varepsilon_{\text{residual}} = \sum_{\mathbf{r}} \varepsilon_{\mathbf{r}}(t_{\text{exp}})$ .

Since  $0 \leq \varepsilon_{\mathbf{r}}(t_{\text{exp}}) < \Theta$  for each pixel, we have

$$\Theta N \leq N_{\text{total}}^{\text{ph}} < \Theta N + \Theta |A|, \quad (124)$$

where  $|A|$  is the number of pixels. The quantity  $\Theta N$  is a lower-biased proxy for the realized total photon count. We therefore use  $\hat{N} = \Theta N$  as an estimate in the high-count regime where  $\Theta |A| \ll \bar{N}$ , yielding an approximate CFAR threshold. A conservative alternative is to replace  $\Theta N$  by  $\Theta N + \Theta |A|$ . The practical CFAR criterion becomes:

$$|\mathcal{E}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \cdot \frac{\Theta N + \Theta |A|}{2}. \quad (125)$$

□

**Remark** (Comparison with the photon CFAR detector). In the single-photon case ( $\Theta = 1$ ), the spike count equals the photon count ( $N = N_{\text{total}}^{\text{ph}}$ ) and the criterion Eq. 125 becomes

$$|\mathcal{E}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \cdot \frac{N}{2}, \quad (126)$$

which matches Equation 5, after accounting for the  $1/\sqrt{v}$  normalization convention.

**Remark** (Role of the quantization residual in detection). The CFAR analysis above assumed  $\eta_{p_{\mathbf{f}}} \approx 0$ . When this does not hold (e.g., at high temporal frequencies where  $|\eta_{p_{\mathbf{f}}}|$  can be  $O(\Theta |f_t| T)$ ), the quantization residual acts as an additional bounded bias under  $H_0$ . This inflates the effective false alarm rate: the test statistic picks up “energy” from the sawtooth quantization pattern even when the true flux has no content at frequency  $\mathbf{f}$ .

This causes artificial harmonic peaks from nearly-periodic spike trains. To maintain the desired false alarm rate, applying a temporal low-pass filter to the spike probing measurement can suppress the high-frequency quantization artifacts before computing  $\mathcal{E}_{\mathbf{f}}$ .

## D.2 Event Cameras: Spatiotemporal Probing Derivation

In this section, we establish the connection of asynchronous event sensors to spatiotemporal probing of the spectrum of  $\partial_t \log \phi$ . For clarity, we focus on the temporal-only case here and extend to spatiotemporal probing in the next section.

To model events as an estimator for the derivative of the log flux, we adopt the mathematical representation presented in [35].

**Event model (per pixel):** Fix a pixel location  $\mathbf{r} = (x, y)$  and let

$$L_{\mathbf{r}}(t) \triangleq \log \phi(\mathbf{r}, t), \quad (127)$$

denote the log-flux (or log-intensity) over an exposure window  $t \in [0, t_{\text{exp}}]$ . In the standard contrast-threshold model, the event camera outputs an event whenever the change in log-flux reaches a contrast threshold  $C > 0$ . For simplicity, we consider the contrast threshold equal for both ON and OFF events. Equivalently, there exists a piecewise-constant quantized process  $\tilde{L}_{\mathbf{r}}(t)$  such that:

1. **(Quantization error)**  $|L_{\mathbf{r}}(t) - \tilde{L}_{\mathbf{r}}(t)| \leq C$  for all  $t \in [0, t_{\text{exp}}]$ ;
2. **(Jump structure)**  $\tilde{L}_{\mathbf{r}}(t)$  changes only via jumps of size  $\pm C$  at the event times  $\{\tau_k(\mathbf{r})\}_{k=1}^{N(\mathbf{r})}$ :

$$\Delta \tilde{L}_{\mathbf{r}}(\tau_k(\mathbf{r})) \triangleq \tilde{L}_{\mathbf{r}}(\tau_k(\mathbf{r})^+) - \tilde{L}_{\mathbf{r}}(\tau_k(\mathbf{r})^-) = \sigma_k(\mathbf{r}) C, \quad \sigma_k(\mathbf{r}) \in \{+1, -1\}. \quad (128)$$

Define the residual (bounded) error

$$\varepsilon_{\mathbf{r}}(t) \triangleq L_{\mathbf{r}}(t) - \tilde{L}_{\mathbf{r}}(t), \quad \text{so that} \quad |\varepsilon_{\mathbf{r}}(t)| \leq C. \quad (129)$$

## D.2.1 Temporal event probing

**Definition 8** (Event field and probing measurements). For a fixed pixel  $\mathbf{r}$ , define the signed event measure (also denoted as an event field in Equation 7 from [35])

$$dS_{\mathbf{r}}(t) \triangleq \sum_{k=1}^{N(\mathbf{r})} \sigma_k(\mathbf{r}) \delta(t - \tau_k(\mathbf{r})) dt, \quad (130)$$

so that, for any probing function  $p$ ,

$$\int_0^{t_{\text{exp}}} p(t) dS_{\mathbf{r}}(t) = \sum_{k=1}^{N(\mathbf{r})} \sigma_k(\mathbf{r}) p(\tau_k(\mathbf{r})). \quad (131)$$

This is the quantized log-intensity signal with respect to the probing function in Equation 8 from [35].

**Proposition 8** (Event probing of  $\partial_t \log \phi$ ). Assume the contrast-threshold event model above and let  $p$  be deterministic, bounded, and differentiable. Then the signed event probing sum is a linear projection of  $\partial_t L_{\mathbf{r}}(t)$  up to a bounded error:

$$C \sum_{k=1}^{N(\mathbf{r})} \sigma_k(\mathbf{r}) p(\tau_k(\mathbf{r})) = \int_0^{t_{\text{exp}}} p(t) \partial_t L_{\mathbf{r}}(t) dt + \eta_{p, \mathbf{r}}, \quad (132)$$

where the residual term satisfies the deterministic bound

$$|\eta_{p, \mathbf{r}}| \leq C \left( |p(t_{\text{exp}})| + |p(0)| \right) + C \int_0^{t_{\text{exp}}} |p'(t)| dt. \quad (133)$$

This is the analogue to Equation 9 from [35], the main difference being that the quantization error in our case depends on the probing function.

*Proof.* Since  $L_{\mathbf{r}}(t) = \tilde{L}_{\mathbf{r}}(t) + \varepsilon_{\mathbf{r}}(t)$ , we write the Stieltjes integral

$$\int_0^{t_{\text{exp}}} p(t) dL_{\mathbf{r}}(t) = \int_0^{t_{\text{exp}}} p(t) d\tilde{L}_{\mathbf{r}}(t) + \int_0^{t_{\text{exp}}} p(t) d\varepsilon_{\mathbf{r}}(t). \quad (134)$$

Because  $\tilde{L}_{\mathbf{r}}(t)$  is piecewise constant with jumps  $\Delta\tilde{L}_{\mathbf{r}}(\tau_k) = \sigma_k C$ ,

$$\int_0^{t_{\text{exp}}} p(t) d\tilde{L}_{\mathbf{r}}(t) = \sum_{k=1}^{N(\mathbf{r})} p(\tau_k(\mathbf{r})) \Delta\tilde{L}_{\mathbf{r}}(\tau_k(\mathbf{r})) = C \sum_{k=1}^{N(\mathbf{r})} \sigma_k(\mathbf{r}) p(\tau_k(\mathbf{r})). \quad (135)$$

For the residual term, apply integration by parts to the Stieltjes integral (valid since both  $p$  and  $\varepsilon_{\mathbf{r}}$  are bounded and measurable):

$$\int_0^{t_{\text{exp}}} p(t) d\varepsilon_{\mathbf{r}}(t) = p(t_{\text{exp}})\varepsilon_{\mathbf{r}}(t_{\text{exp}}) - p(0)\varepsilon_{\mathbf{r}}(0) - \int_0^{t_{\text{exp}}} \varepsilon_{\mathbf{r}}(t) p'(t) dt. \quad (136)$$

Using  $|\varepsilon_{\mathbf{r}}(t)| \leq C$  gives

$$\left| \int_0^T p(t) d\varepsilon_{\mathbf{r}}(t) \right| \leq C(|p(T)| + |p(0)|) + C \int_0^T |p'(t)| dt. \quad (137)$$

Finally, since  $dL_{\mathbf{r}}(t) = \partial_t L_{\mathbf{r}}(t) dt$  for differentiable  $L_{\mathbf{r}}$ , we have  $\int_0^{t_{\text{exp}}} p(t) dL_{\mathbf{r}}(t) = \int_0^{t_{\text{exp}}} p(t) \partial_t L_{\mathbf{r}}(t) dt$ . Combining with Eq. 134–Eq. 135 yields the event probing sum Eq. (132) by setting  $\eta_{p,\mathbf{r}} \triangleq \int_0^{t_{\text{exp}}} p(t) d\varepsilon_{\mathbf{r}}(t)$  as the bound Eq. 133.  $\square$

## D.2.2 Spatiotemporal event probing (proof of Equation 9)

In this section, we extend the temporal analysis to the spatiotemporal domain and derive Equation 9 of the main paper.

**Equation 9.** The Fourier probing measurements for event cameras are

$$\mathcal{E}_{\mathbf{f}} = \sum_{(x_i, y_i, t_i) \in \mathcal{P}} \sigma_i p_{\mathbf{f}}(x_i, y_i, t_i). \quad (9)$$

**Event stream over the sensor:** Let  $A$  denote the set of pixel locations (or a continuous sensor domain with pixel area  $|A|$ ). The event camera outputs a set of events

$$\mathcal{V} = \{(\mathbf{r}_i, t_i, \sigma_i)\}_{i=1}^N, \quad \mathbf{r}_i = (x_i, y_i) \in A, \quad t_i \in [0, t_{\text{exp}}], \quad \sigma_i \in \{+1, -1\}. \quad (138)$$

Define the global signed event measure on  $A \times [0, T]$ :

$$dS(\mathbf{r}, t) \triangleq \sum_{i=1}^N \sigma_i \delta(\mathbf{r} - \mathbf{r}_i) \delta(t - t_i) d\mathbf{r} dt, \quad (139)$$

so that for any probing function  $p(\mathbf{r}, t)$ ,

$$\int_A \int_0^{t_{\text{exp}}} p(\mathbf{r}, t) dS(\mathbf{r}, t) = \sum_{i=1}^N \sigma_i p(\mathbf{r}_i, t_i). \quad (140)$$

Substituting  $p(\mathbf{r}, t) = p_{\mathbf{f}}(\mathbf{x}) = e^{-j2\pi \mathbf{f}^T \mathbf{x}}$  in Eq. (140) yields Equation 9 of the main paper.

**Proposition 9** (Spatiotemporal event probing). Assume the per-pixel contrast-threshold model holds at every  $\mathbf{r} \in A$ , i.e.,  $L_{\mathbf{r}}(t) = \log \phi(\mathbf{r}, t)$  admits a quantized approximation  $\tilde{L}_{\mathbf{r}}(t)$  with  $|L_{\mathbf{r}}(t) - \tilde{L}_{\mathbf{r}}(t)| \leq C$  and jumps of size  $\pm C$  at event times with polarities. Let  $p(\mathbf{r}, t)$  be bounded and differentiable in  $t$  for each  $\mathbf{r}$ . Then

$$C \sum_{i=1}^N \sigma_i p(\mathbf{r}_i, t_i) = \int_A \int_0^{t_{\text{exp}}} p(\mathbf{r}, t) \partial_t \log \phi(\mathbf{r}, t) dt d\mathbf{r} + \eta_p, \quad (141)$$

where the residual term  $\eta_p$  is bounded by the spatial integral of the per-pixel bounds:

$$|\eta_p| \leq \int_A \left[ C(|p(\mathbf{r}, T)| + |p(\mathbf{r}, 0)|) + C \int_0^{t_{\text{exp}}} |\partial_t p(\mathbf{r}, t)| dt \right] d\mathbf{r}. \quad (142)$$

*Proof.* Apply Proposition 8 independently to each pixel  $\mathbf{r}$  with probing function  $t \mapsto p(\mathbf{r}, t)$ , then sum the resulting identities over  $\mathbf{r} \in A$  (or integrate over  $\Omega$  in the continuous-domain view). The left-hand side becomes  $\sum_i \sigma_i p(\mathbf{r}_i, t_i)$  as in Eq. 140, while the right-hand side becomes the space–time integral of  $p(\mathbf{r}, t) \partial_t \log \phi(\mathbf{r}, t)$  plus the sum (or integral) of residual terms. The bound Eq. 142 follows by summing the per-pixel bound Eq. 133.  $\square$

### D.2.3 Simplified noise model for events

While Proposition 9 provides an upper bound for the error of the probing measurements, it does not represent a stochastic model to derive a CFAR detector. We introduce background activity (BA) as the main source of shot noise, where circuit junction leakage and temperature periodically produce ON events [26].

To keep the model simple, we make the assumption that BA events originate from a homogeneous Poisson point process, where the events are biased towards ON events. This is a common assumption used in event denoising methods [10, 22] and event camera simulators [15].

**Definition 9** (Biased background activity model). Given a pixel  $\mathbf{r} \in A$ , the background activity model assumes the following:

- (i) Background events arrive at  $\mathbf{r}$  with homogeneous Poisson process rate  $\lambda_{\text{BA}}$  events/second. For simplicity, we assume all pixels adopt the same rate.
- (ii) Each background activity event has polarity  $\sigma_i = \pm 1$ , where  $\sigma_i = +1$  with probability  $q > 1/2$  due to ON bias, and  $\sigma_i = -1$  with probability  $1 - q$ , independently drawn.
- (iii) Background activity events are independent of other events and between pixels, which is a standard simplification that neglects refractory-period interactions between the two streams.

We denote the events coming from real signal as  $\mathcal{V}_{\text{sig}}$ , and those coming from background activity as  $\mathcal{V}_{\text{BA}}$ , thus we have that

$$\mathcal{V} = \mathcal{V}_{\text{sig}} \cup \mathcal{V}_{\text{BA}}. \quad (143)$$

The total expected number of background activity events over the sensor and observation window is

$$\mathbb{E}[\mathcal{V}_{\text{BA}}] = \bar{N}_{\text{BA}} = \int_A \int_0^{t_{\text{exp}}} \lambda_{\text{BA}} dt d\mathbf{r} = |A| \lambda_{\text{BA}} t_{\text{exp}}. \quad (144)$$

**Corollary 4.** For any BA event  $(\mathbf{r}, t, \sigma) \in \mathcal{V}_{\text{BA}}$ , its polarity  $\sigma \in \{+1, -1\}$  satisfies:

$$\mathbb{E}[\sigma] = 2q - 1 \triangleq b, \quad (145)$$

$$\mathbb{E}[\sigma^2] = 1, \quad (146)$$

$$\text{Var}(\sigma) = 1 - b^2 = 4q(1 - q). \quad (147)$$

We refer to  $b \in (-1, 1]$  as the *polarity bias*. Note that  $\mathbb{E}[\sigma^2] = 1$  identically since  $\sigma^2 = 1$  for all events regardless of polarity.

### D.2.4 Fourier probing: null distribution

Here, we derive the distribution of the null hypothesis of the probing measurements  $H_0$  for the purposes of CFAR detection.

To make the CFAR analysis tractable, we work in the regime where the quantization error  $\eta_p$  is small. That is, since the bound of  $\eta_p$  scales linearly with frequency  $\mathbf{f}$  (due to the  $\int_0^{t_{\text{exp}}} |\partial_t p(\mathbf{r}, t)| dt$  in Eq. (142)), the quantization error is small if the scene does not vary too quickly. Naturally, this makes faster scenes harder to reconstruct.

Using the Fourier basis as our probing function, denote our polarity-weighted probing measurements as

$$\mathcal{E}_{\mathbf{f}} = \mathcal{E}_{\mathbf{f}}^{\text{sig}} + \mathcal{E}_{\mathbf{f}}^{\text{BA}}, \quad (148)$$

such that  $\mathcal{E}_{\mathbf{f}}^{\text{sig}} = C \sum_{i=1}^{N_{\text{sig}}} \sigma_i e^{-j2\pi \mathbf{f}^\top (\mathbf{r}_i, t_i)}$  and  $\mathcal{E}_{\mathbf{f}}^{\text{BA}} = C \sum_{i=1}^{N_{\text{BA}}} \sigma_i e^{-j2\pi \mathbf{f}^\top (\mathbf{r}_i, t_i)}$ .

Note that from Proposition 9,  $\mathcal{E}_{\mathbf{f}}^{\text{sig}}$  is deterministic up to a bounded error.

Under the null hypothesis  $H_0$  that  $\mathbf{f}$  is not a component of  $\partial_t \log \phi(\mathbf{r}, t)$ , we expect  $\mathcal{E}_{\mathbf{f}}^{\text{sig}} = 0$ , therefore  $\mathcal{E}_{\mathbf{f}} = \mathcal{E}_{\mathbf{f}}^{\text{BA}}$ .

**Proposition 10** (Mean and variance of  $P_{\mathbf{f}}^{\text{BA}}$  under  $H_0$ ). Under the background activity model (Definition 9) with symmetric contrast threshold  $C$ , the background activity probing measurement  $\mathcal{E}_{\mathbf{f}}^{\text{BA}}$  can be modeled as a functional of a marked Poisson point process [24], since  $\mathcal{V}_{\text{BA}}$  itself is a homogeneous Poisson point process. For  $\mathbf{f} \neq 0$ , its mean and variance are as follows:

$$\mathbb{E}[\mathcal{E}_{\mathbf{f}}^{\text{BA}}] = 0 \quad (149)$$

$$\text{Var}(\mathcal{E}_{\mathbf{f}}^{\text{BA}}) = C^2 \bar{N}_{\text{BA}}. \quad (150)$$

We restrict frequencies to values that are integer multiples of the grid resolution ( $1/w, 1/h, 1/t_{\text{exp}}$  for  $f_x, f_y, f_t$  respectively), so that the Fourier bases are orthonormal and integrate over full periods over the domain.

*Proof of Proposition 10.* By Campbell's theorem (Lemma 3), the mean and variance of a sum over a Poisson process are:

$$\mathbb{E}[\mathcal{E}_{\mathbf{f}}^{\text{BA}}] = \lambda_{\text{BA}} \sum_{\sigma \in \{+1, -1\}} g(\sigma) \int_A \int_0^{t_{\text{exp}}} C \sigma e^{-j2\pi \mathbf{f}^\top(\mathbf{r}, t)} dt d\mathbf{r}, \quad (151)$$

$$\text{Var}(\mathcal{E}_{\mathbf{f}}^{\text{BA}}) = \lambda_{\text{BA}} \sum_{\sigma \in \{+1, -1\}} g(\sigma) \int_A \int_0^{t_{\text{exp}}} |C \sigma e^{-j2\pi \mathbf{f}^\top(\mathbf{r}, t)}|^2 dt d\mathbf{r}, \quad (152)$$

where  $g(+1) = q$  and  $g(-1) = 1 - q$ .

**Mean.** Evaluating Eq. 151, and using the fact that polarity  $\sigma$  is independent of the event location:

$$\mathbb{E}[\mathcal{E}_{\mathbf{f}}^{\text{BA}}] = C \lambda_{\text{BA}} \underbrace{\mathbb{E}[\sigma]}_{=b} \int_A \int_0^{t_{\text{exp}}} e^{-j2\pi \mathbf{f}^\top(\mathbf{r}, t)} dt d\mathbf{r}. \quad (153)$$

We restrict to frequencies that are integer multiples of the grid resolution ( $1/w, 1/h, 1/t_{\text{exp}}$ ), so that the Fourier basis functions complete full periods over the observation volume. As a consequence, the Fourier basis is orthonormal and  $\int_A \int_0^{t_{\text{exp}}} e^{-j2\pi \mathbf{f}^\top(\mathbf{r}, t)} dt d\mathbf{r} = 0$ , therefore

$$\mathbb{E}[\mathcal{E}_{\mathbf{f}}^{\text{BA}}] = 0, \quad \mathbf{f} \neq 0. \quad (154)$$

For the DC component  $\mathbf{f} = 0$ , the mean is  $Cb \lambda_{\text{BA}} |A| t_{\text{exp}} = Cb \bar{N}_{\text{BA}}$ .

**Variance.** Evaluating Eq. 152:

$$\begin{aligned} \text{Var}(\mathcal{E}_{\mathbf{f}}^{\text{BA}}) &= C^2 \lambda_{\text{BA}} \underbrace{\sum_{\sigma \in \{+1, -1\}} g(\sigma) \sigma^2}_{= \mathbb{E}[\sigma^2] = 1} \int_A \int_0^{t_{\text{exp}}} \underbrace{|e^{-j2\pi \mathbf{f}^\top(\mathbf{r}, t)}|^2}_{=1} dt d\mathbf{r} \\ &= C^2 \lambda_{\text{BA}} |A| t_{\text{exp}} = C^2 \bar{N}_{\text{BA}}. \end{aligned} \quad (155)$$

□

**Proposition 11** (Covariance of real and imaginary parts). Let  $\mathcal{E}_{\mathbf{f}}^{\text{BA}} = X_{\mathbf{f}} + jY_{\mathbf{f}}$  such that  $X_{\mathbf{f}} = C \sum_{i \in \mathcal{V}_{\text{BA}}} \sigma_i \cos(2\pi \mathbf{f}^\top(\mathbf{r}_i, t_i))$  and  $Y_{\mathbf{f}} = -C \sum_{i \in \mathcal{V}_{\text{BA}}} \sigma_i \sin(2\pi \mathbf{f}^\top(\mathbf{r}_i, t_i))$ , we have that

$$\Sigma = \text{Var}(X_{\mathbf{f}}) = \text{Var}(Y_{\mathbf{f}}) = \frac{C^2 \bar{N}_{\text{BA}}}{2}, \quad \text{Cov}(X_{\mathbf{f}}, Y_{\mathbf{f}}) = 0. \quad (156)$$

*Proof of Proposition 11.* Campbell's theorem applied separately gives:

$$\text{Var}(X_{\mathbf{f}}) = C^2 \lambda_{\text{BA}} \int_A \int_0^{t_{\text{exp}}} \cos^2(2\pi \mathbf{f}^\top(\mathbf{r}, t)) dt d\mathbf{r}, \quad (157)$$

$$\text{Var}(Y_{\mathbf{f}}) = C^2 \lambda_{\text{BA}} \int_A \int_0^{t_{\text{exp}}} \sin^2(2\pi \mathbf{f}^\top(\mathbf{r}, t)) dt d\mathbf{r}, \quad (158)$$

$$\text{Cov}(X_{\mathbf{f}}, Y_{\mathbf{f}}) = -C^2 \lambda_{\text{BA}} \int_A \int_0^{t_{\text{exp}}} \cos(2\pi \mathbf{f}^\top(\mathbf{r}, t)) \sin(2\pi \mathbf{f}^\top(\mathbf{r}, t)) dt d\mathbf{r}. \quad (159)$$

We restrict to frequencies that are orthonormal and integrate over full periods in the domain. As a consequence, the integrals of the Fourier basis evaluate to

$$\int_A \int_0^{t_{\text{exp}}} \cos^2(2\pi \mathbf{f}^\top(\mathbf{r}, t)) dt d\mathbf{r} = \frac{|A| t_{\text{exp}}}{2}, \quad (160)$$

$$\int_A \int_0^{t_{\text{exp}}} \sin^2(2\pi \mathbf{f}^\top(\mathbf{r}, t)) dt d\mathbf{r} = \frac{|A| t_{\text{exp}}}{2}, \quad (161)$$

$$\int_A \int_0^{t_{\text{exp}}} \cos(2\pi \mathbf{f}^\top(\mathbf{r}, t)) \sin(2\pi \mathbf{f}^\top(\mathbf{r}, t)) dt d\mathbf{r} = 0, \quad (162)$$

using the following properties:  $\cos^2 \theta = \frac{1}{2}(1 + \cos 2\theta)$ ,  $\sin^2 \theta = \frac{1}{2}(1 - \cos 2\theta)$ ,  $\cos \theta \sin \theta = \frac{1}{2} \sin 2\theta$ , and the double-frequency terms integrating to zero on the grid. Substituting and evaluating  $\bar{N}_{\text{BA}} = \lambda_{\text{BA}} |A| t_{\text{exp}}$ :

$$\text{Var}(X_{\mathbf{f}}) = \text{Var}(Y_{\mathbf{f}}) = \frac{C^2 \bar{N}_{\text{BA}}}{2}, \quad \text{Cov}(X_{\mathbf{f}}, Y_{\mathbf{f}}) = 0. \quad (163)$$

□

**Proposition 12** (Asymptotic normality). For large  $\bar{N}_{\text{BA}}$ , the CLT for Poisson integrals gives:

$$\mathcal{E}_{\mathbf{f}}^{\text{BA}} \stackrel{d}{\approx} \mathcal{CN}(0, C^2 \bar{N}_{\text{BA}}), \quad \mathbf{f} \neq \mathbf{0}, \quad (164)$$

i.e.,  $X_{\mathbf{f}}$  and  $Y_{\mathbf{f}}$  are approximately independent each with  $\mathcal{N}(0, C^2 \bar{N}_{\text{BA}}/2) = \mathcal{N}(0, \Sigma)$ .

*Proof of Proposition 12.* Conditioned on  $N_{\text{BA}} = n$ , the quantity

$$\mathcal{E}_{\mathbf{f}}^{\text{BA}} = C \sum_{i=1}^n \sigma_i e^{-j2\pi \mathbf{f}^\top(\mathbf{r}_i, t_i)}$$

is a sum of  $n$  i.i.d. bounded complex random variables. For nonzero Fourier frequencies, these summands have mean zero, and their real and imaginary parts have equal variance. Therefore, for sufficiently large  $n$ , the multivariate central limit theorem implies

$$\mathcal{E}_{\mathbf{f}}^{\text{BA}} | (N_{\text{BA}} = n) \stackrel{d}{\approx} \mathcal{CN}(0, C^2 n).$$

Taking expectation over  $N_{\text{BA}}$  and replacing  $n$  with its expectation  $\bar{N}_{\text{BA}}$  gives the unconditional approximation in Eq. (164). □

With the null distribution established, we now derive the CFAR detector.

## D.2.5 CFAR detector for event cameras

**Corollary 5** (Normalized probing energy under  $H_0$ ). Define the normalized probing energy at frequency  $\mathbf{f} \neq \mathbf{0}$ :

$$\tilde{\mathcal{E}}_{\mathbf{f}} \triangleq \frac{X_{\mathbf{f}}^2}{\Sigma} + \frac{Y_{\mathbf{f}}^2}{\Sigma} = \frac{|\mathcal{E}_{\mathbf{f}}|^2}{\Sigma}, \quad \Sigma = \frac{C^2 \bar{N}_{\text{BA}}}{2} \quad (165)$$

Under  $H_0(\mathbf{f})$ , where no signal is present at frequency  $\mathbf{f}$ , we have  $\mathcal{E}_{\mathbf{f}} = \mathcal{E}_{\mathbf{f}}^{\text{BA}}$ , and thus  $\tilde{\mathcal{E}}_{\mathbf{f}} \sim \chi_2^2$  (chi-squared with 2 degrees of freedom) asymptotically.

**Proposition 13** (CFAR detector for event camera Fourier probing). Given a false alarm probability  $\alpha$ , we detect spatiotemporal frequency  $\mathbf{f} \neq \mathbf{0}$  (reject  $H_0$ ) if and only if

$$|\mathcal{E}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \frac{C^2 \bar{N}_{\text{BA}}}{2}. \quad (166)$$

However, since  $\bar{N}_{\text{BA}}$  is not directly observable, we instead use the total event count  $N$  as a conservative upper bound ( $\bar{N}_{\text{BA}} \leq N$  under most cases).

$$|\mathcal{E}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \frac{C^2 N}{2}. \quad (167)$$

This more conservative bound produces fewer detections, where the actual false alarm rate is at most  $\alpha$ .

*Proof.* Under  $H_0$ ,  $\tilde{\mathcal{E}}_{\mathbf{f}} = |\mathcal{E}_{\mathbf{f}}|^2 / \Sigma \sim \chi_2^2$ . Thus, evaluating with respect to the CDF of  $\chi_2^2$  gives

$$|\mathcal{E}_{\mathbf{f}}|^2 \geq \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \Sigma = \text{CDF}_{\chi_2^2}^{-1}(1 - \alpha) \frac{C^2 \bar{N}_{\text{BA}}}{2}. \quad (168)$$

□

## E Photon-Location Prediction and Global Flicker: an Intuitive Analysis for bit2bit

In our experiments, bit2bit [28] often preserves fine spatial detail, yet can attenuate weak scene-wide illumination flicker. We do not view this as a fundamental limitation of the method. Rather, we offer the following intuition for why a photon-location prediction objective, when trained on short spatiotemporal crops, may provide a stronger and more consistent learning signal for spatial structure than for weak global temporal modulation.

**Setup:** Consider a simple global-flicker model

$$\phi(x, y, t) = A(t) \lambda(x, y), \quad (169)$$

where  $\lambda(x, y)$  is a fixed spatial pattern and  $A(t)$  is a time-varying gain shared by all pixels. Under this model, flicker changes primarily *when* photons arrive, while leaving their relative *spatial distribution* essentially unchanged.

**A simple distributional view:** Let  $p(i, t)$  denote the normalized distribution of held-out target detections over a crop, where  $i$  indexes spatial location and  $t$  indexes time bin. Under Eq. 169, this distribution factorizes as

$$p(i, t) = p_t s_i, \quad (170)$$

where  $p_t$  is the temporal marginal induced by  $A(t)$  and  $s_i$  is the normalized spatial pattern induced by  $\lambda(x, y)$ .

For intuition, suppose the model’s predicted distribution can be viewed in the form

$$q_\theta(i, t) = q_\theta(t) q_\theta(i | t). \quad (171)$$

Then the expected negative log-likelihood of  $M$  held-out detections can be written as

$$\mathbb{E}[\mathcal{L}(\theta)] = \text{const} + M \text{KL}(p_t \| q_\theta(t)) + M \sum_t p_t \text{KL}(s \| q_\theta(\cdot | t)). \quad (172)$$

Here,  $M = \sum_{t=1}^T \sum_{i=1}^N w_{i,t}$  is the number of held-out target detections in the crop,  $w_{i,t} \in \{0, 1\}$  indicates whether voxel  $(i, t)$  contains a held-out target detection, and  $\text{KL}(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$  denotes the Kullback–Leibler divergence. Equation (172) should be read as an *interpretive* decomposition rather than an exact statement about the implementation of bit2bit [28]. Its purpose is to highlight that, under global flicker, the temporal signal is carried only by the low-dimensional marginal  $p_t$ , whereas the spatial term reflects the full spatial arrangement of detections within each crop.

**Why weak flicker may be de-emphasized:** Under global flicker, the spatial pattern  $s_i$  is approximately time-invariant, so every held-out detection contributes consistent supervision for spatial prediction. By contrast, the only cue for flicker is the variation of the 1D temporal marginal  $p_t$  across the crop. If that within-crop variation is weak, then a temporally smoothed prediction  $q_\theta(t)$  may incur only a modest penalty, even when the model fits spatial structure well.

This effect is expected to be strongest when the crop duration is short relative to the flicker period. For example, if

$$A(t) = 1 + \epsilon \sin(2\pi f t), \quad (173)$$

then over a crop of duration  $\Delta_{\text{win}} = T\Delta t$ , the maximum change in gain is

$$\Delta A_{\text{max}} = 2\epsilon |\sin(\pi f \Delta_{\text{win}})|. \quad (174)$$

For  $f = 120$  Hz, frame rate 100 kHz, and  $T = 32$  (this is bit2bit’s default temporal crop [28]), we have  $\Delta_{\text{win}} = 0.32$  ms, which gives

$$\Delta A_{\text{max}} \approx 0.048 \quad \text{when } \epsilon = 0.2. \quad (175)$$

Thus, even a 20% sinusoidal flicker produces less than 5% variation within such a crop, so the temporal histogram of detections can appear nearly flat at crop scale.

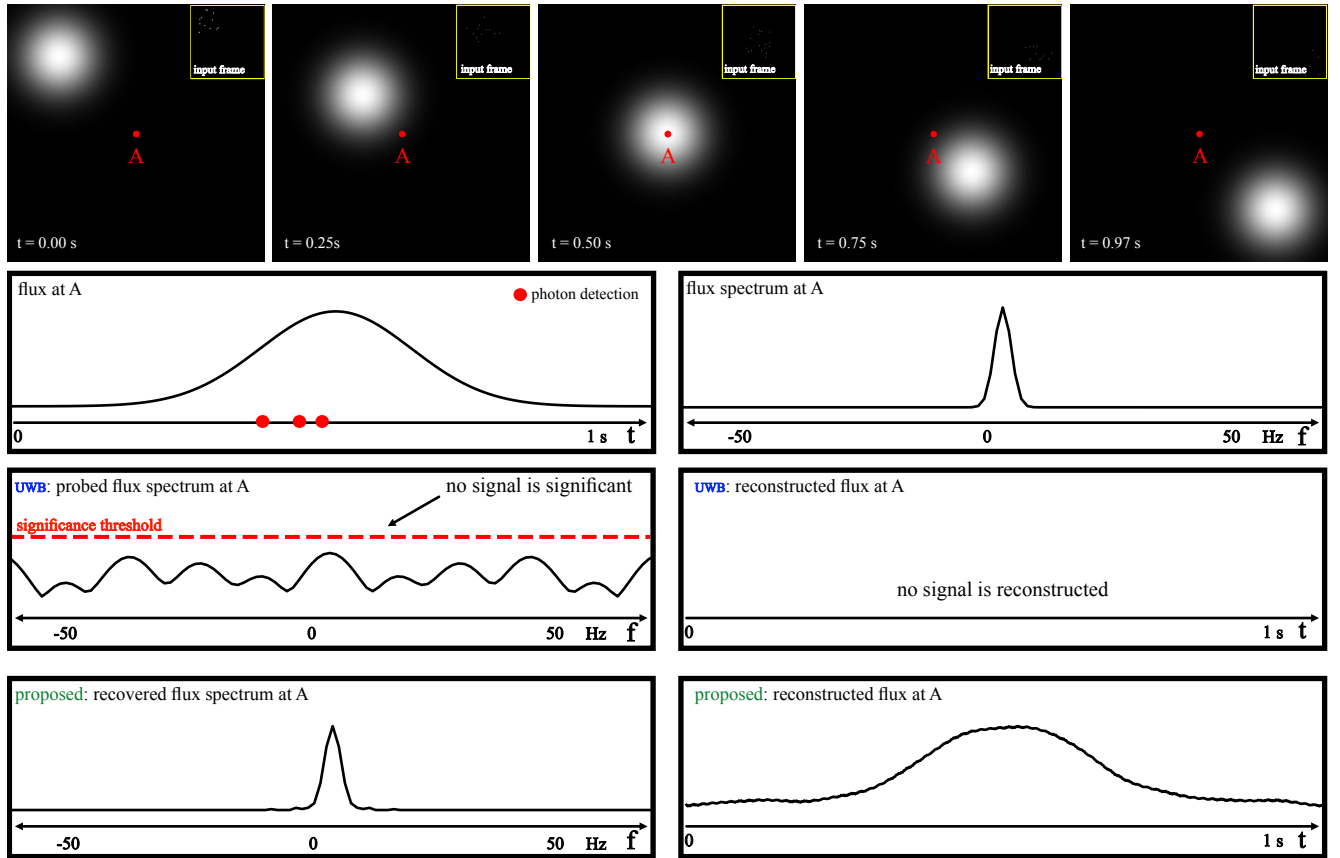
**Interpretation:** The above analysis does not imply that bit2bit cannot represent flicker. Rather, it suggests that, in the short-crop low-photon regime considered in this work, the objective may offer a stronger and more stable incentive to refine spatial photon placement than to preserve weak crop-wide temporal modulation. This provides one possible explanation for why bit2bit can produce visually sharp reconstructions while still under-emphasizing subtle global flicker in our experiments (see Fig. 1, top row, and Fig. 4, rows 1 and 3 of the main paper; the temporal smoothing is most clearly visible in the videos on the supplemental webpage).

## F Limitations of Per-Pixel Temporal Analysis Under Motion

**Why per-pixel temporal methods fail under motion:** When a scene moves, photon detections are spread across many pixels rather than concentrated at a fixed location. At any single pixel, the object is visible only during the brief interval in which it passes through, so only a small fraction of the total photon budget contributes to the local temporal signal. Under limited photon counts, this sparsity can be severe enough that the per-pixel temporal spectrum never exceeds the CFAR threshold [41] at any frequency, causing the reconstruction to fail entirely.

**Illustration:** Figure F.1 demonstrates this effect using a translating Gaussian blob. Although the blob is smooth and its temporal profile at a fixed pixel is a simple pulse (Row 2, left), only three photons are detected at pixel A over the full acquisition. The resulting per-pixel spectrum (Row 3, left) is noise-dominated and remains below the significance threshold at all frequencies, so UWB [41] reconstructs no signal at A.

**Global spatiotemporal probing:** The key observation is that per-pixel methods discard the spatial dimension entirely: each pixel is processed independently using only its own detections. By contrast, spatiotemporal probing computes 3D Fourier coefficients by summing over all photon arrivals across both space and time, aggregating evidence from the entire sensor rather than a single pixel. Even though pixel A receives only three photons, the global photon count is orders of magnitude



**Figure F.1: Failure of pixelwise temporal probing under motion and extreme photon sparsity.** **Row 1:** Spatiotemporal flux frames of a translating Gaussian blob, with corresponding binary input photon frames shown in the insets. As the blob passes through point A, the local flux varies smoothly in time, while the observed binary frames contain only a few isolated photon arrivals. **Row 2:** Ground truth temporal flux at pixel A, along with the three detected photons (left), and the ground truth flux spectrum (right). **Row 3:** UWB [41] applied at pixel A: the per-pixel temporal spectrum remains below the significance threshold at all frequencies, so no component is detected and the temporal signal is not reconstructed. This is due to extreme photon sparsity at pixel A and the absence of strong periodicity in the local signal. **Row 4:** By aggregating photons across space and time rather than relying solely on detections at pixel A, our method recovers a clean spectrum and reconstructs a temporal signal that closely matches the ground-truth flux.

larger. Our method recovers a clean spectrum (Row 4, left) and reconstructs a temporal signal at A that closely matches the ground truth (Row 4, right).

## G Implementation Details

### G.1 Single-photon videography

We discuss implementation details of performing spatiotemporal flux probing on different single-photon cameras.

#### G.1.1 Frequency selection and spatiotemporal normalization

The spatiotemporal frequency grid is determined by the physical extent of the acquisition window. The frequency resolution in each dimension is  $1/w$ ,  $1/h$ , and  $1/t_{\text{exp}}$  in  $x$ ,  $y$ , and  $t$  respectively, and we probe integer multiples of these resolutions up to the Nyquist limits set by the sensor’s spatial and temporal sampling rates.

To ensure orthonormality of the Fourier basis with respect to the integral over the physical spatiotemporal volume, we normalize each probing function by  $1/\sqrt{v}$ , where  $v = w \cdot h \cdot t_{\text{exp}}$  is the total physical volume of the window. The normalized basis functions are then

$$p_{\mathbf{f}}(\mathbf{x}) = \frac{1}{\sqrt{v}} e^{-j2\pi\mathbf{f}^\top \mathbf{x}}, \quad (176)$$

which satisfy

$$\langle p_{\mathbf{f}}, p_{\mathbf{f}'} \rangle = \begin{cases} 0 & \mathbf{f} \neq \mathbf{f}' \\ 1 & \mathbf{f} = \mathbf{f}' \end{cases} \quad (177)$$

under the physical-domain integral. Crucially, this normalization ensures that the reconstructed flux  $\hat{\phi}(\mathbf{x})$  is expressed in the correct physical units.

The computation of the Fourier probing measurements depends on the sensor type. For quanta sensors [34] the binary readout produces regularly sampled frames, so the probing measurements can be computed efficiently via the fast Fourier transform (FFT) [5]. For SPAD arrays [31], photon arrivals are irregularly sampled in time, so we compute the probing measurements via direct sum of the terms, which evaluates the Fourier sum at the irregular photon timestamps without binning into discrete frames.

#### G.1.2 Selection of the CFAR significance level $\alpha$

The significance level  $\alpha$  controls the tradeoff between sensitivity to weak frequencies and robustness to noise. A larger  $\alpha$  lowers the detection threshold, admitting more frequencies and recovering finer spatiotemporal detail, but increases the expected number of false alarms. A smaller  $\alpha$  raises the threshold, suppressing spurious detections at the cost of potentially missing weaker signal components.

A natural choice is  $\alpha = 1/|\mathcal{F}|$ , where  $|\mathcal{F}|$  is the number of probed frequencies. Under the CFAR model this corresponds to controlling the expected number of false detections across the entire frequency grid to approximately one. In practice, however, we found this choice to be overly conservative: the resulting threshold suppresses many high-frequency components and produces over-smoothed reconstructions.

In all experiments we therefore use  $\alpha = 10^{-4}$ , which provides a good balance between recovering fine spatiotemporal structure and limiting spurious frequency detections.

#### G.1.3 Short-time Fourier transform for quanta sensors

The spatiotemporal flux probing framework places no constraints on how the probing functions partition or aggregate the spatiotemporal volume—it holds for any bounded, measurable  $p$ , including globally supported ones. In principle, one can probe the entire acquisition at once, and for SPAD arrays this is what we do: the smaller sensor format and shorter acquisition windows make global Fourier probing practical.

For quanta sensors, however, larger spatial extent and longer acquisition windows make a single global Fourier model both less practical and less well matched to the nonstationary structure of natural scenes. Sharp spatiotemporal boundaries produce broadband spectral leakage under a global Fourier basis, and hard frequency thresholding then introduces Gibbs-like ringing artifacts in the reconstruction [11, 13]. This is a well-known limitation of the Fourier basis under nonstationarity, not of the probing theory itself. We therefore adopt the short-time Fourier transform (STFT) [1, 6] with Hann windows [13] and 75% overlap. Our default window size is  $100 \times 100$  pixels  $\times$  10,000 frames. This trades global frequency resolution for

spatiotemporal localization, and empirically produces more coherent reconstructions of scenes with spatially and temporally varying structure.

Even with windowing, the spatiotemporal context of each STFT block far exceeds that of existing methods. Each window aggregates  $100 \times 100 \times 10,000 = 10^8$  spatiotemporal samples, providing roughly  $300\times$  more temporal context than bit2bit [28] and  $100\times$  more spatial context than per-pixel temporal analysis like UWB<sup>1</sup>. The STFT therefore preserves the core advantage of spatiotemporal probing—pooling photon evidence across a large spatiotemporal volume—while adapting to the local structure of natural scenes. In other words, the STFT is not a theoretical necessity but a modeling choice: natural scenes are spatiotemporally nonstationary, and local windows assume only that the flux is locally well-approximated by a superposition of sinusoids—a much weaker assumption than global stationarity.

### G.1.4 Windowed CFAR detection and reconstruction

When using the short-time Fourier transform for quanta image sensors, the global CFAR detector and reconstruction must be adapted to operate within each local window. Define  $w(\mathbf{x})$  as the window function. Within each spatiotemporal window  $W_k$  of volume  $v_k = |W_k|$ , frequency detection proceeds exactly as in the global setting (Equation 5), but with the probing function  $p_{\mathbf{f},k}(\mathbf{x}) = \frac{1}{\sqrt{v_k}} w(\mathbf{x} - \mathbf{x}_k) e^{-j2\pi \mathbf{f}^\top \mathbf{x}}$ . Approximating its covariance matrix as the diagonal matrix in Corollary 1, we obtain

$$\text{Var}(\text{Re}[\mathcal{E}_{\mathbf{f},k}]) \approx \text{Var}(\text{Im}[\mathcal{E}_{\mathbf{f},k}]) \approx \frac{\langle w(\mathbf{x} - \mathbf{x}_k)^2, \phi \rangle}{2v_k} \approx \frac{|\mathcal{P}|_{w,k}}{2v_k}, \quad (178)$$

where the window-weighted photon count

$$|\mathcal{P}|_{w,k} = \sum_{\mathbf{x} \in \mathcal{P}} w(\mathbf{x} - \mathbf{x}_k)^2 \quad (179)$$

is an empirical estimate of the effective photon count within window  $k$ , with each photon weighted by the squared window value at its location. We detect frequency  $\mathbf{f}$  within window  $k$  with significance level  $\alpha$  if

$$|\mathcal{E}_{\mathbf{f},k}|^2 \geq \text{CDF}_{\chi^2}^{-1}(1 - \alpha) \frac{|\mathcal{P}|_{w,k}}{2v_k}. \quad (180)$$

This is a windowed CFAR detector: the threshold adapts to the local noise level within each window, as estimated from the window-weighted photon count  $|\mathcal{P}|_{w,k}$ , ensuring a consistent false alarm rate across all windows regardless of local photon density.

**Weighted overlap-add:** To reconstruct a continuous-time flux estimate from overlapping windowed measurements, we use a weighted overlap-add (WOLA) procedure [6]. Let  $\hat{\phi}_k(\mathbf{x})$  denote the local flux reconstruction within window  $k$ , obtained by summing the detected Fourier coefficients. The spatiotemporal flux  $\hat{\phi}(\mathbf{x})$  is then reconstructed by the window-weighted sum

$$\hat{\phi}(\mathbf{x}) = \frac{\sum_k w(\mathbf{x} - \mathbf{x}_k) \hat{\phi}_k(\mathbf{x})}{\sum_k w(\mathbf{x} - \mathbf{x}_k)^2}, \quad (181)$$

where the denominator normalizes for the fact that the window is applied twice: once during analysis (via the windowed probing function) and once during synthesis (via the numerator weighting). Defining  $w^2$  as the Hann window with 75% overlap, the squared window sum  $\sum_k w(\mathbf{x} - \mathbf{x}_k)^2$  is approximately constant over the interior of  $\Omega$  [13], so the reconstruction simplifies to a uniform weighted combination of overlapping windows.

## G.2 Velocity-selective videography

In this section, we describe the implementation details of velocity detection and velocity-selective videography, corresponding to Sections 3.2 and 4 of the main paper, respectively.

### G.2.1 Velocity detection hyperparameters

Our velocity detection algorithm has the following hyperparameters:

<sup>1</sup>Quanta image sensors output  $\sim 100\text{k}$  binary frames per second, so UWB [41] probes each pixel independently using all  $\sim 10^5$  temporal samples. Our STFT window aggregates  $100 \times 100 \times 10,000 = 10^8$  spatiotemporal samples—roughly  $1000\times$  more total measurements per probed region, at the cost of using  $10\times$  fewer temporal frames per pixel.

1. velocity plane thickness  $\epsilon$ , fixed at 0.5 unless specified differently;
2. velocity search range  $(v_{\min}, v_{\max}, N_v)$ ;
3. neighborhood window  $w$ ;
4. neighborhood guard  $g$ , fixed at 3 unless specified differently;
5. CFAR rate  $\alpha_{\text{vel}}$  fixed at 0.001 unless differently specified.

The plane thickness  $\epsilon$  determines how much spectral mass is counted as supporting a given motion plane. Larger values improve robustness to discretization, noise, and slight model mismatch, but also reduce selectivity by allowing more off-plane coefficients to contribute. Smaller values yield sharper energy concentration but can become brittle when the spectral support is broadened by finite sampling, noise, or deviations from constant-velocity motion.

The number of sampled velocities per axis,  $N_v$ , controls the discretization of the search space: a finer grid improves velocity resolution but increases the cost of plane aggregation quadratically.

The effects of neighborhood geometry and CFAR rate  $\alpha_{\text{vel}}$  can be summarized as contributing to the control of overall false detection rate.

## G.2.2 Velocity detection with single-photon sensors

Given a candidate velocity  $\mathbf{v} = [v_x, v_y]^\top$ , we measure how much spectral energy is consistent with constant motion at that velocity, and detect significant velocities using a nonparametric CFAR test. Starting from the power spectrum  $\mathcal{E}^2(f_x, f_y, f_t)$  of the input volume, detection proceeds in three steps:

1. For each candidate velocity, compute the signed distance of every frequency-domain point to the corresponding velocity plane,

$$\xi(f_x, f_y, f_t; \mathbf{v}) = f_t + v_x f_x + v_y f_y, \quad (182)$$

and sum the power within thickness  $\epsilon$  of the plane to obtain the energy map,

$$E(\mathbf{v}) = \sum_{(f_x, f_y, f_t)} \mathcal{E}^2(f_x, f_y, f_t) \mathbf{1}[|\xi(f_x, f_y, f_t; \mathbf{v})| \leq \epsilon]. \quad (183)$$

Velocities are swept over a uniform grid of  $N_v$  values per axis, normalized as described in Section G.2.3 to match the discrete Fourier grid.

2. For each cell  $i$  in the energy map, build a training set  $\mathcal{T}_i$  from the neighborhood of half-width  $w$  around  $i$ , excluding an inner guard region of half-width  $g < w$ .
3. Compute the rank statistic, the number of training cells exceeded by the CUT,

$$r_i = \sum_{e \in \mathcal{T}_i} \mathbf{1}[E_i \geq e]. \quad (184)$$

Under  $H_0$ ,  $r_i$  is uniform on  $\{0, \dots, |\mathcal{T}_i|\}$ , as proved in Equation 8. A detection is declared at  $i$  when

$$r_i \geq \lceil (1 - \alpha_{\text{vel}})(|\mathcal{T}_i| + 1) \rceil. \quad (185)$$

For a spatiotemporal crop of size  $w \times h \times N_t$  and a velocity grid of size  $N_v \times N_v$ , computing the energy map requires comparing every spectral sample against every candidate plane, giving  $O(N_t h w N_v^2)$  time. The subsequent CFAR test operates only on the  $N_v \times N_v$  energy map, adding  $O(N_v^2 w^2)$  time using a sliding-window implementation.

## G.2.3 Velocity-selective videography with single-photon sensors

After selecting a velocity  $\mathbf{v} = [v_x, v_y]^\top$ , we form a velocity-selective video by compensating the corresponding motion in the Fourier domain. Under constant velocity, the 3D spectrum of the scene is supported on a plane through the origin (Equation 6 in the main paper). The goal is to de-shear this plane so that it aligns with  $f_t = 0$ , effectively freezing the selected motion.

The compensation is a three-step procedure. Starting from the 3D spectrum  $\Phi(f_x, f_y, f_t)$  of the input volume:

1. Apply an inverse FFT along the temporal axis, producing a mixed representation in  $[f_x, f_y, t]^\top$ .
2. Multiply by the phase ramp

$$e^{i 2\pi t (v_x^{\text{sc}} f_x + v_y^{\text{sc}} f_y)} \quad (186)$$

which shifts each spatial frequency by a temporal offset proportional to the velocity.

3. Apply a forward FFT along time to obtain the de-sheared spectrum, from which the compensated video is reconstructed.

Reconstruction is then performed in a sliding temporal window over the photon stream to generate video frames. Although related ideas have been explored in conventional videography [40] and single-photon imaging [37], our framework performs velocity focusing at ultra-low light levels and without requiring manual specification of the motion direction or magnitude as in previous work [37].

The scaled velocities  $v_x^{\text{sc}}$  and  $v_y^{\text{sc}}$  account for the fact that the candidate velocities are specified in pixels per frame, while the phase ramp operates on the discrete Fourier grid. The rescaling matches the two coordinate systems:

$$v_x^{\text{sc}} = v_x \frac{\max(f_t) - \min(f_t)}{\max(f_x) - \min(f_x)}, \quad v_y^{\text{sc}} = v_y \frac{\max(f_t) - \min(f_t)}{\max(f_y) - \min(f_y)}. \quad (187)$$

For a spatiotemporal crop of size  $w \times h \times N_t$ , each compensated video requires one inverse and one forward FFT along time plus one element-wise multiplication, giving  $O(N_t h w \log N_t)$  time and  $O(N_t h w)$  memory. All operations are dense tensor computations and run efficiently on a GPU.

### G.3 Spike camera videography

Spike cameras produce spikes at regularly sampled frames at 40k fps [16], so we take the STFT approach from single-photon videography, instead using the conservative CFAR bound defined in Eq. (125).

To suppress harmonic artificial frequencies produced from constant flux spike trains, we apply a low-pass filter to the probing measurements by dropping frequencies  $\mathbf{f}$  with large temporal frequencies  $f_t$ .

### G.4 Event camera videography

We discuss implementation details of reconstructing log-flux from event cameras, both from pure integration using only events, and a basic fusion method using a Kalman filter [20].

#### G.4.1 Log-flux estimation from only events

Using the spatiotemporal event probing measurements described in Proposition 9, we apply CFAR detection using the conservative bound in Eq. (167) with  $\alpha = 10^{-4}$ . This yields a noisy estimate of  $\partial_t \log \phi(\mathbf{r}, t)$ , which we can then integrate over  $t_{\text{exp}}$  for an estimate of  $\log \phi(\mathbf{r}, t)$  up to a constant.

Since our reconstructed function takes the form of

$$\widehat{\partial_t \log \phi(\mathbf{r}, t)} = C \sum_{\mathbf{f} \in \mathcal{F}_{\text{det}}} \mathcal{E}_{\mathbf{f}} e^{j 2\pi \mathbf{f}^\top [\mathbf{r}, t]}, \quad (188)$$

integration over time can be performed by dividing the spectrum by  $j 2\pi f_t$  for  $f_t \neq 0$ :

$$\widehat{\log \phi(\mathbf{r}, t)} = C \sum_{\mathbf{f} \in \mathcal{F}_{\text{det}}, f_t \neq 0} \frac{1}{j 2\pi f_t} \mathcal{E}_{\mathbf{f}} e^{j 2\pi \mathbf{f}^\top [\mathbf{r}, t]}. \quad (189)$$

We discard all temporal DC components ( $f_t = 0$ ), which are generally dominated by bias and drift [35], and apply a high-pass filter in  $f_t$  to suppress additional low-frequency artifacts in the reconstructed scene.

### G.4.2 Kalman filter with intensity frames

The event data in [32] was recorded with a DAVIS240C camera [18], which also records intensity images at approximately 20 fps. This can be combined with the estimated log-flux derivative  $\partial_t \widehat{\log \phi(\mathbf{r}, t)}$  via the Kalman filter [20].

We model the log-flux  $x(\mathbf{r}, t) = \log \phi(\mathbf{r}, t)$  as a scalar state per pixel. The process model integrates the event-derived derivative estimate as a control input over a timestep  $\Delta t$ :

$$x(\mathbf{r}, t) = x(\mathbf{r}, t - \Delta t) + \Delta t \cdot \partial_t \widehat{\log \phi(\mathbf{r}, t)} + w, \quad w \sim \mathcal{N}(0, Q), \quad (190)$$

and the intensity measurement model treats each APS frame as a direct noisy observation of the log-flux:

$$z(\mathbf{r}) = \log I(\mathbf{r}) = x(\mathbf{r}) + v, \quad v \sim \mathcal{N}(0, R). \quad (191)$$

The same filter applies identically in the Fourier domain: replacing  $x(\mathbf{r}, t)$  with  $\hat{x}(f, t)$  and  $z(\mathbf{r})$  with  $\hat{z}(f)$ , the process and measurement models retain their linear-Gaussian structure, since the Fourier transform is a linear operation. In our experiments,  $Q$  and  $R$  were set empirically without formal tuning. A principled calibration of these noise parameters is left to future work.

## H Experiments

We use the SPAD512 camera from Pi Imaging to capture the new datasets introduced with this work. This camera produces binary frames, which are single-bit images recording a 1 at a pixel location if it receives one or more photons during a frame exposure. Unless otherwise stated, each binary frame has a temporal resolution (exposure) of 10 microseconds, and frames are taken at a regular rate of 97.1 kfps for a total of 100000 frames. Scene illumination is controlled to ensure that photon detections are unlikely to be affected by SPAD dead time. In other words, 1s in the binary frames are unlikely to correspond to more than a single photon arrival. Under this operating condition, treating detections as individual photon arrivals is effectively equivalent to sampling from the underlying 3D inhomogeneous Poisson process governed by the photon arrival rate at the sensor. To further ensure this regime, we additionally employ ND filters and adjust the aperture of the focusing lens when necessary.

### Scene specific details:

- **Balloon Popping with 31 kHz LED (Fig. 1).** The scene contains two sources of illumination flicker: ceiling lights powered by the AC mains at approximately 120 Hz, and an additional light source flickering at 31 kHz. An ND filter with optical density  $OD=0.7$  was used for this capture. The average detection rate was 2458 photons per pixel per second (ppps), varying from 14625 ppps in bright regions to 34 ppps in dark regions.
- **Blender (Fig. 4).** The illumination flicker in this scene is also caused by AC-powered lighting, resulting in a natural flicker at approximately 120 Hz. An ND filter with optical density  $OD = 0.7$  is used for this scene. The average detection rate was 5977 photons per pixel per second (ppps), varying from 98885 ppps in bright regions to 75 ppps in dark regions.
- **Static Balloon with high-speed light (Fig. 4).** The 31 kHz flickering light source is generated using an LED–resistor circuit driven by a function generator with a sinusoidally varying voltage. Ambient lighting with 120 Hz AC flicker is also present in the scene. An ND filter with optical density  $OD = 0.7$  is used. The average detection rate was 5933 photons per pixel per second (ppps), varying from 14937 ppps in bright regions to 496 ppps in dark regions.
- **Dropping Ball (Fig. 4).** A static ball is released from a fixed height and falls with minimal rotational motion. An ND filter with optical density  $OD = 0.3$  is used for this scene. The average detection rate was 3278 photons per pixel per second (ppps), varying from 18105 ppps in bright regions to 45 ppps in dark regions.
- **Balloon Popping with 20 kHz LED (supplemental webpage).** We captured an additional scene similar to Fig. 1, in which a 20 kHz flickering LED is directed at the balloon to make the flicker more visible. The background illumination flicker in this scene originates from ceiling lights powered by the AC mains, producing a flicker at approximately 120 Hz. An ND filter with optical density  $OD=0.7$  was used for this capture. The average detection rate was 3653 photons per pixel per second (ppps), varying from 33937 ppps in bright regions to 59 ppps in dark regions.

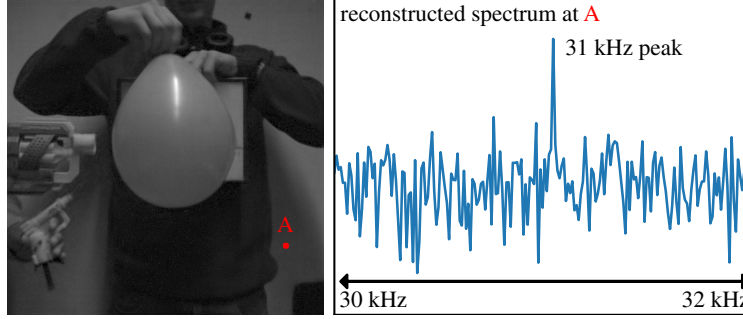
**Thinning while maintaining dark counts:** To simulate a lower photon flux, we perform thinning of the photon detections by independently applying Bernoulli sampling to each detected event, retaining it with probability  $p$ . It is well known that thinning an inhomogeneous Poisson process in this manner produces another inhomogeneous Poisson process whose rate is scaled by  $p$  relative to the original process [36]. Note that uniformly thinning all timestamps by a constant factor does not change the SNR, since both signal photons and dark counts are reduced proportionally. To vary the SNR, we thin the original set of captured timestamps while maintaining the same level of dark counts.

We estimate the dark count rate of the SPAD512 single-photon camera by capturing binary frames with the lens cap on. After removing hot pixels following [28], we compute the average dark count rate as the total number of detected photons divided by the total acquisition time and number of pixels, yielding 25 counts per second (cps) for our camera. We then thin the original timestamps by a desired factor  $p$  and add timestamps from a homogeneous Poisson process with rate  $(1 - p) \times 25$  cps. This reduces the signal by approximately factor  $p$  while maintaining a constant dark count rate of 25 cps, effectively lowering the SNR.

**Hot pixel correction:** Hot pixel correction is performed through device calibration and then using statistical off-board correction as implemented in the `spadtools` [27] library.

### H.1 Single-photon videography

For all scenes captured with the SPAD512 single-photon camera (binary frame quanta sensor), we use the STFT with  $100 \times 100 \times 10,000$  windows, Hann windowing, and 75% overlap, as described in Section G.1.3. Within each window, we probe spatial frequencies  $f_x$  and  $f_y$  up to 256 cycles and temporal frequency  $f_t$  up to 50 kHz. We set  $\alpha = 10^{-4}$  and reconstruct the spatiotemporal flux from the detected frequencies. Since the reconstructed flux is a continuous function, it can be sampled at



**Figure H.2: Recovered 31 kHz temporal flicker in Figure 1.** Our method successfully recovers the faint 31 kHz LED flicker, alongside the 120 Hz ceiling light flicker and the scene dynamics shown in Figure 1.

any desired frame rate up to the Nyquist limit set by the sensor’s timestamp resolution. We perform gamma correction to all videos with  $\gamma = 2.2$ . The supplemental webpage shows videos rendered at various frame rates for the scenes in Figure 1 and Figure 4 (top) of the main paper.

**Remark.** Frequency detection is essential to the reconstruction. Applying the Fourier transform and inverting it without frequency selection effectively reproduces the original binary sequence, placing approximately a Dirac impulse at every photon location in space and time.

### H.1.1 Additional details for Balloon popping with 31 kHz LED (Figure 1 scene)

Our method is able to recover the 31 kHz modulated LED in Figure 1 simultaneously with the projectiles, 120 Hz ceiling light, and balloon rupture. We show the recovered temporal spectrum at a pixel in Fig. H.2.

### H.1.2 Additional details for ultra-wideband videography (Figure 4, row 5 scene)

Following UWB [41], we probe temporal frequencies corresponding to the known harmonics of the fan (108 Hz) and the MHz laser (20 MHz), since an exhaustive 3D Fourier transform up to GHz temporal frequencies is computationally infeasible. For each selected  $f_t$ , we probe all spatial frequencies and apply CFAR detection to reject outlier spatial frequencies.

### H.1.3 Runtime for single-photon videography

**Computational complexity.** On quanta data, the binary frames form a regularly sampled volume of  $w \times h$  pixels over  $T$  time bins spanning a physical exposure  $t_{\text{exp}}$ , giving  $N = whT$  total samples. Global 3D FFT+CFAR+IFFT costs  $O(N \log N)$ . With the STFT, let  $K$  denote the number of spatiotemporal windows, each of size  $N_k = w_k h_k T_k$  samples; the total cost is  $O(K N_k \log N_k)$ , where  $K$  grows with overlap and shrinking window size, which incurs higher cost than the global FFT.

**Runtime.** All reconstructions run on a single NVIDIA V100 GPU (32 GB VRAM) and require no training. For SPAD512 scenes, reconstructing 10,000 consecutive frames takes under 15 minutes, while reconstructing 1,000 frames sparsely sampled across all 100,000 quanta frames takes approximately one hour. For reference, bit2bit requires approximately 8 hours of training on 40,000 binary frames before inference can begin.

## H.2 Velocity-selective videography on the Figure 1 scene

We demonstrate velocity-selective videography on this real scene containing motions of varying speed, contrast, and spatial extent. The velocity search grid spans  $300 \times 300$  hypotheses (half-width 150 in each direction), yielding 90,000 candidate velocities. The neighborhood window is  $w = 150$  and the significance level is  $\alpha_{\text{vel}} = 10^{-4}$ . For each event of interest, we select the velocity range and reference frame to match the speed and timing of the target motion.

Figure H.3 shows four representative motions from the same scene. From top to bottom, the targets are: the white bullet, the balloon, the gray bullet, and the ECCV logo. Each row contains four panels: (i) the velocity-space detection map with detected velocities marked, (ii) the sum of binary frames over the selected temporal interval, (iii) the velocity-selective reconstruction for the detected motion, and (iv) the QBP reconstruction over the same interval. Orange boxes highlight the region corresponding to the target motion. Full video sequences are available on the supplemental webpage.

**Velocity detection:** We report the velocity detection results for each target motion below:

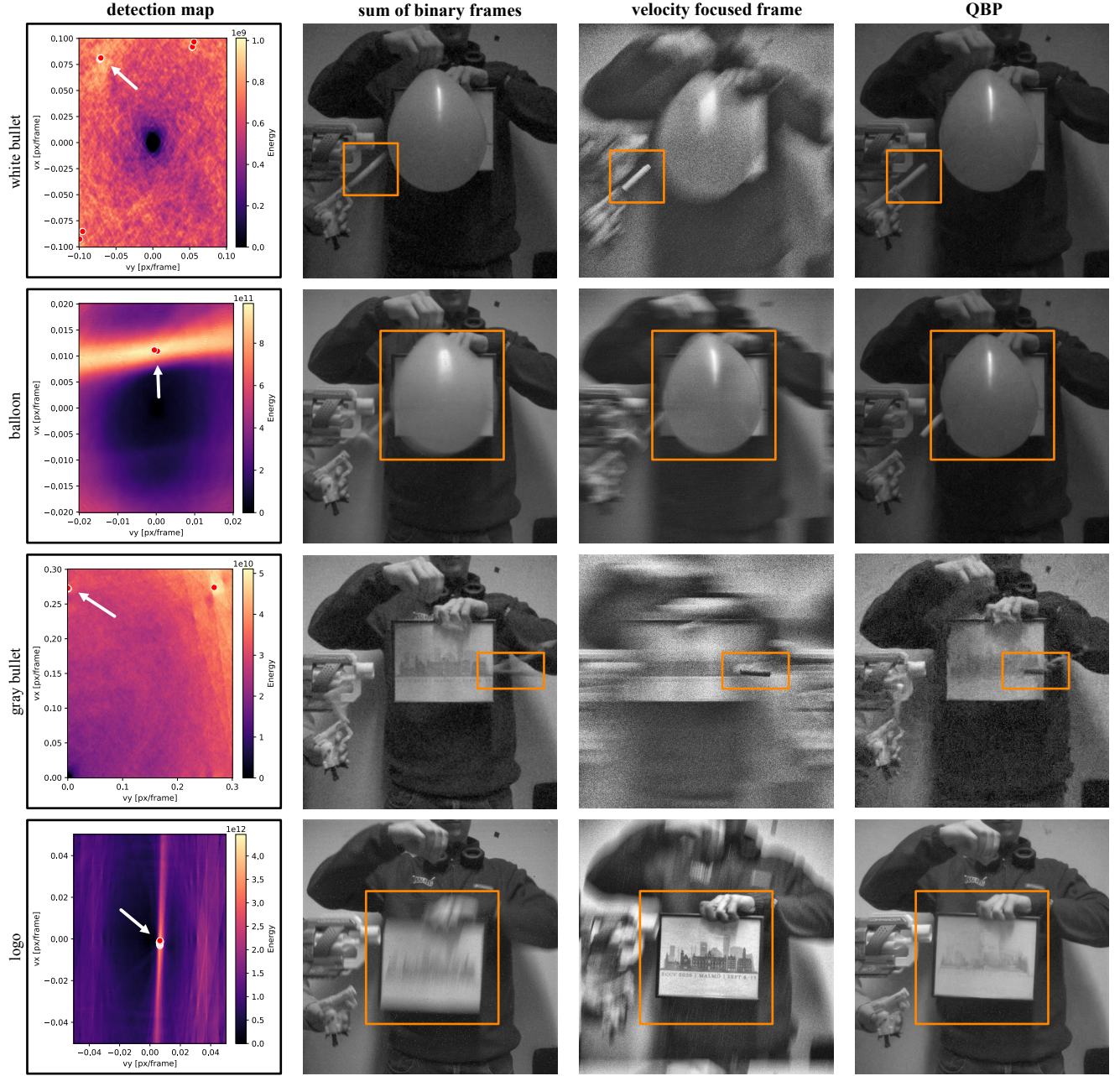
- **White bullet:** 9 detections, 4 inliers. Detected velocity:  $(0.076, -0.073)$  px/frame.
- **Balloon:** 3 detections, 2 inliers. Detected velocity:  $(0.011, -0.001)$  px/frame. The balloon undergoes nonlinear motion (when hit by the white bullet), but at the frame rate of the SPAD512 ( $\sim 100$  kfps) its trajectory is well approximated as locally linear. The successful detection suggests that our velocity detection method can extend beyond strictly constant-velocity scenes.
- **Gray bullet:** 3 detections, 2 inliers. Detected velocity:  $(0.272, 0)$  px/frame.
- **ECCV logo:** 8 detections, all inliers. Detected velocity:  $(-0.001, -0.007)$  px/frame.

**Observations:** Overall, the velocity detector recovers accurate motion hypotheses across a range of velocities and object sizes, and the resulting reconstructions successfully isolate the target motion from the rest of the scene. Interestingly, the velocity-selective reconstruction isolates the target motion even when the summed binary frames contain substantial overlap from other scene content. This is most evident for the white (Fig. H.3, row 1) and gray bullets (Fig. H.3, row 3), where the moving object is difficult to distinguish in the summed frames but becomes clearly separated after velocity focusing. The balloon (Fig. H.3, row 2) demonstrates that the method also applies to larger, slower motions, whose broader velocity-space signatures yield smoother reconstructions. The ECCV logo (Fig. H.3, row 4) produces a particularly coherent detection, with all hypotheses classified as inliers, indicating a strong and well-isolated motion signature. Compared with QBP [29], our method selectively reconstructs the velocity of interest while suppressing other scene content, producing sharper images of the target object.

### H.3 Videography with other asynchronous sensors (Figure 5)

We apply event probing to the datasets `shapes_6dof` (top row, left) and `office_spiral` (top row, right) from [32]. We note that choosing to apply the Kalman filter in image spatial or Fourier domain affects the quality of the reconstruction depending on the scene. As a result, we apply Kalman filtering in the spatial domain for `shapes_6dof`, and in the Fourier domain in `office_spiral`.

We reconstruct the following spike camera videos from [42]: `balloon`, `rotation2x`, and `car-100kmh`. As we discuss in the main paper, pixels with constant flux produce nearly periodic spike trains, which introduce spurious harmonic peaks at high frequencies. To suppress these artifacts, we apply a temporal low-pass filter whose cutoff depends on scene brightness: in darker regions, the inter-spike interval is longer and the spurious harmonics appear at lower frequencies, requiring a more aggressive cutoff.



**Figure H.3: Velocity-selective videography on the Fig. 1 scene.** We apply velocity detection to four moving objects: the white bullet, balloon, gray bullet, and ECCV logo (rows top to bottom). Each row shows: the velocity-space energy map with detected velocities (red circles); the sum of binary frames over the selected temporal interval; the velocity-selective reconstruction focused on the detected velocity (annotated with the white arrow); and the QBP reconstruction over the same interval. Orange boxes highlight the target object. The velocity-selective reconstruction isolates each moving object from the scene, whereas the summed frames and QBP mix all motions together. Full video sequences are available on the supplemental webpage.

# I Simulations

## I.1 Probing measurement normality

### I.1.1 Central limit theorem validity

The probing measurements  $\mathcal{E}_p$  tend towards a normal distribution as the number of observed photons increases (Proposition 2), with the rate of convergence quantified via the Berry–Esseen theorem [2, 9], which bounds the worst-case deviation between the CDF of a standardized sum of  $n$  i.i.d. random variables and the standard normal CDF by  $C\rho/(\sigma^3\sqrt{n})$ , where  $\sigma^2$  is the variance,  $\rho = \mathbb{E}[|X - \mu|^3]$  is the third absolute central moment, and  $C > 0$  is a universal constant. The ratio  $\rho/\sigma^3$  characterizes how many samples are needed for a given accuracy, with heavier-tailed distributions requiring more.

For quanta image sensors in the Poisson regime, the CLT sample size is the number of photons contributing to a spatiotemporal probe. For a typical STFT window of  $100 \times 100 \times 10,000$  space-time bins at  $10^{-4}$  ppp, this gives  $\sim 10^4$  photons per probe, which as we show empirically in Section I.1.2 is well above the threshold for normality to hold.

In the high- $q$  Bernoulli regime, each space-time bin produces a Bernoulli random variable with success probability  $q$ . The Berry–Esseen factor for a Bernoulli( $q$ ) variable is

$$\frac{\rho}{\sigma^3} = \frac{(1-q)^3 \cdot q + q^3 \cdot (1-q)}{(q(1-q))^{3/2}} = \frac{q^2 + (1-q)^2}{\sqrt{q(1-q)}}, \quad (192)$$

which grows rapidly as  $q \rightarrow 1$ : at  $q = 0.999$ , this factor is  $\sim 32\times$  larger than at  $q = 0.5$ , meaning convergence to normality requires  $\sim 10^3\times$  more samples for comparable accuracy. However, for a probe with  $n = 10^8$  space-time bins at  $q = 0.999$ , the expected number of nondetections is  $(1-q)n \approx 10^5$ , which are the primary drivers of variance in the probing sum. Despite the  $\sim 10^3\times$  higher sample requirement relative to  $q = 0.5$ , the large probe support ensures the effective number of nondetections remains well into the regime where the CLT approximation holds.

### I.1.2 Empirical validation

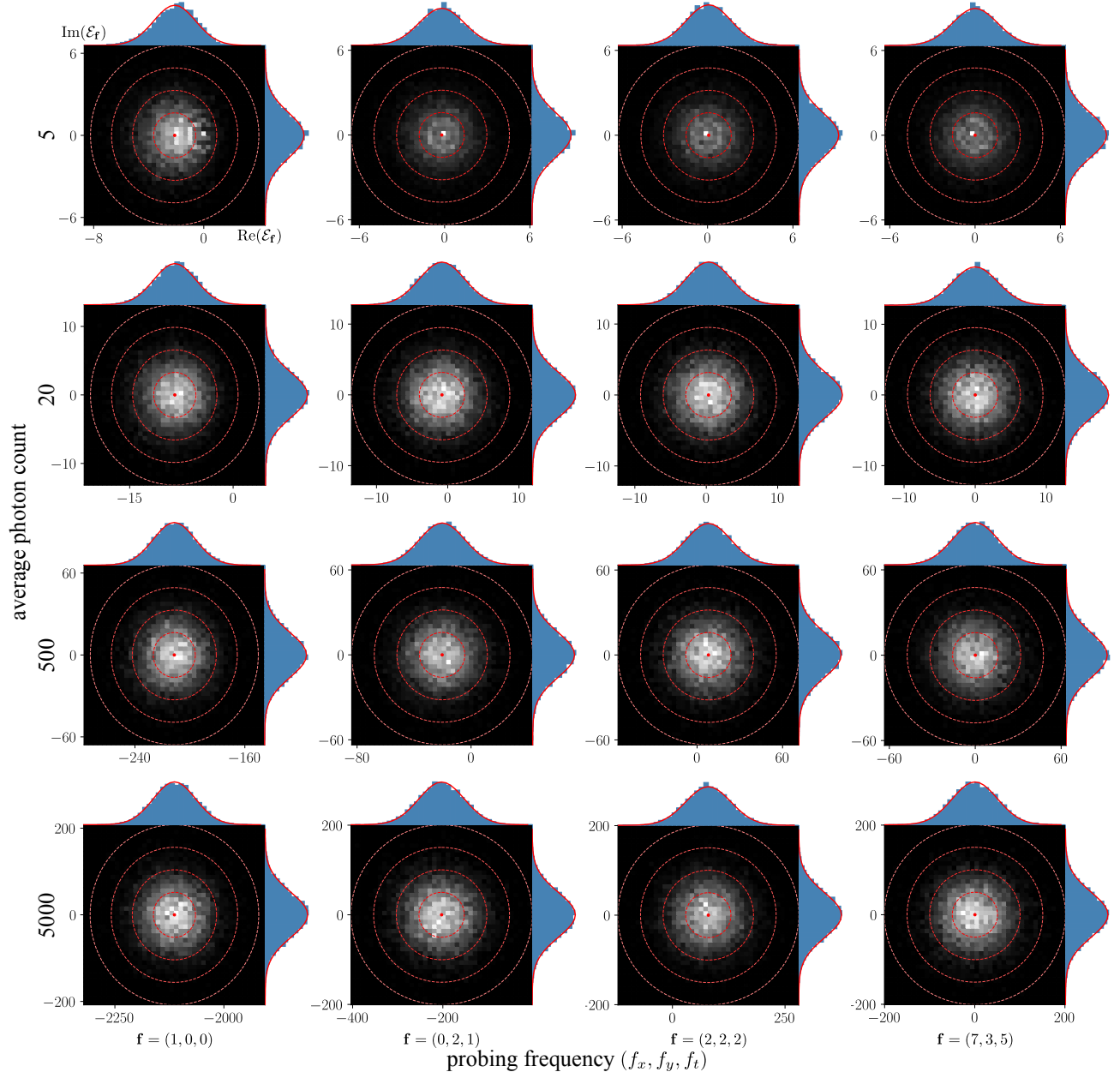
We validate the normality approximation empirically through simulations in Fig. I.4, where we probe photons realized from the same Gaussian blob flux function used in Fig. F.1 via the Fourier probing basis. The flux is scaled to different amplitudes to simulate different light levels, such that the expected number of photons for each simulation trial is 5, 20, 500, and 5,000 respectively. For each light level, 10,000 trials are run. For each trial, we compute the probing measurements from the realized photons at selected frequencies and estimate the distribution of each  $\mathcal{E}_f$ .

The empirical probing measurement histograms in Fig. I.4 are overlaid with ground-truth Gaussian contours computed from Equation 3, centered at the ground-truth mean, where each ring represents one standard deviation. For as few as 20 photons, both the joint and marginal distributions are already well approximated by a complex Gaussian. This is well within the photon counts observed in our experiments: even our darkest thinned datasets, with an STFT window of  $100 \times 100$  pixels  $\times$  10,000 frames, operate at light levels significantly above this threshold. Furthermore, the contours are approximately circular across all selected frequencies, consistent with the diagonal covariance approximation of Equation 3; note however that this reflects the spectral properties of the Gaussian blob rather than natural video statistics, for which the diagonal approximation is instead justified by the power law decay of natural video power spectra (Corollary 1).

## I.2 CFAR threshold validity

The CFAR detection threshold in Equation 5 uses the observed photon count  $|\mathcal{P}|$  as a proxy for the unknown flux integral  $\langle 1, \phi \rangle$ . Since  $|\mathcal{P}|$  is itself a Poisson random variable with mean  $\langle 1, \phi \rangle$  and standard deviation  $\sqrt{|\mathcal{P}|}$ , the threshold is subject to shot noise. We evaluate the robustness of this approximation in Fig. I.5 by comparing the set of frequencies recovered under the observed-count threshold against the oracle threshold derived from the true flux integral, measured via mean intersection-over-union (IoU) of the detected frequency masks. We also evaluate thresholds at  $|\mathcal{P}| \pm 3\sqrt{|\mathcal{P}|}$ , representing the extremes of a  $3\sigma$  fluctuation in the observed count, to characterize the worst-case sensitivity of the detector to photon count noise.

At our experimental operating regime of  $\geq 10^{-3}$  ppp ( $\geq 10^5$  photons), the observed-count threshold achieves a mean IoU close to 1 relative to the oracle bound, confirming that  $|\mathcal{P}|$  is a reliable proxy for  $\langle 1, \phi \rangle$  at practical flux levels. It also performs better on average compared to the extreme  $\pm 3\sqrt{|\mathcal{P}|}$  bounds, suggesting that underestimating or overestimating the bound is not necessary.

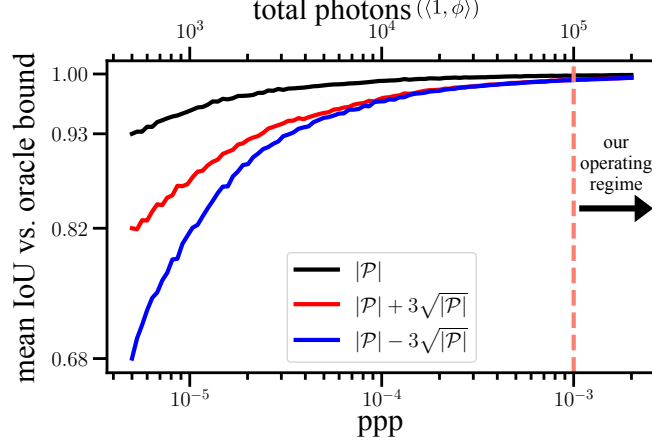


**Figure I.4: Distribution of Fourier probing measurements at varying light levels.** Each cell shows the empirical joint distribution of the real and imaginary parts of  $\mathcal{E}_f$  over 10,000 trials, with marginal histograms (blue) and ground-truth normal curves (red) shown along the top and right axes. Ground-truth Gaussian contours (red), spaced one standard deviation apart, are overlaid on each 2D histogram. Columns correspond to selected probe frequencies assuming the domain is a unit spatiotemporal cube; rows correspond to expected photon counts of 5, 20, 500, and 5,000. The approximation is already accurate at 20 photons, well below the photon counts observed in our experiments.

### I.3 Quantitative reconstruction error

#### I.3.1 Simulating photon arrivals from high-speed videos

We simulate binary frames captured by a single photon camera from high-speed videos captured by conventional cameras using a procedure similar to [38]:



**Figure I.5: CFAR threshold robustness.** Mean IoU of recovered frequencies relative to the oracle CFAR threshold (derived from the true flux integral  $\langle 1, \phi \rangle$ ), evaluated over 1,000 trials per light level. The observed-count threshold  $|\mathcal{P}|$  (black) and  $\pm 3\sigma$  bounds  $|\mathcal{P}| \pm 3\sqrt{|\mathcal{P}|}$  (red/blue) are compared against the oracle. The bottom axis shows photons-per-pixel (ppp) for an STFT window of  $100 \times 100$  pixels  $\times$  10,000 frames; the top axis shows the corresponding total photon count. Our experimental operating regime ( $\geq 10^{-3}$  ppp, indicated by the arrow) lies in the region where all three thresholds converge to near-perfect agreement with the oracle.

1. We collect high-speed videos originally filmed with a Phantom camera [14], at 1 kfps. The videos are downloaded at  $1920 \times 1080$ , and are downsampled to  $512 \times 288$ . The gamma-corrected videos are converted to linear intensity (exponent 2.2).
2. To increase temporal resolution, we first interpolate the videos by  $16\times$  using RIFE [17], yielding a nominal frame rate of 16 kfps. We then linearly interpolate 7 additional frames between consecutive frames, producing a final temporal resolution of approximately 128 kfps.
3. Photon timestamps are generated at each pixel by thinning an inhomogeneous Poisson process whose rate is assumed proportional to the interpolated intensity signal. We treat the normalized intensity values in the range  $[0, 1]$  as representing an approximately constant photon arrival rate within each pixel. The resulting binary videos finally are quantized to 10-microsecond resolution.

### I.3.2 Simulated evaluation

Table 1 evaluates our video reconstruction method across 6 simulated scenes at three photon levels, covering a broad range of flux conditions. We compare against UWB [41], QNN [39] and bit2bit [28], the current state-of-the-art in photon-limited video reconstruction. QBP [29] is computationally prohibitive for practical frame-by-frame video synthesis<sup>2</sup>, so we omit it from the quantitative comparison. We do, however, include QBP in the supplemental videos and in the velocity-selective videography comparison (Section H.2), given its consistently strong reconstruction quality and widespread use as a reference method [3].

As expected, reconstruction quality degrades as the photon level decreases for all methods. Our method achieves the highest PSNR on the large majority of scene and light-level combinations. Perceptual metrics (LPIPS and SSIM) are more competitive: our method leads on most scenes, but bit2bit achieves stronger perceptual scores on scenes with smooth textures and liquid motion (*e.g.*, strawberry milk, grass waterfall), where its learned denoiser can hallucinate plausible detail that improves perceptual similarity without necessarily improving fidelity. Full video reconstructions for all scenes and methods are available on the supplemental webpage.

<sup>2</sup>Using the implementation of [29], processing the raw  $130k \times 512 \times 512$  binary volume and rendering 300 output frames requires 2–3 days on a single CPU.

**Table 1:** Quantitative comparisons across six simulated scenes at three photons per pixel (ppp) levels. Higher PSNR, SSIM, and lower LPIPS indicate better image reconstruction quality. The best result among the four methods for each scene and ppp level is indicated in **red**.

| Metric             | Method   | ppp   | Scene        |              |                |               |                |               |
|--------------------|----------|-------|--------------|--------------|----------------|---------------|----------------|---------------|
|                    |          |       | coffeebeans  | fireblast    | grasswaterfall | shoewaterstep | strawberrymilk | wineglassfall |
| PSNR $\uparrow$    | proposed | 0.1   | <b>29.18</b> | 33.21        | <b>26.53</b>   | <b>31.26</b>  | <b>30.75</b>   | <b>33.14</b>  |
|                    |          | 0.01  | <b>27.98</b> | 31.74        | 26.71          | <b>27.97</b>  | <b>28.45</b>   | <b>28.74</b>  |
|                    |          | 0.002 | <b>25.38</b> | 28.52        | <b>24.77</b>   | <b>22.19</b>  | <b>20.08</b>   | <b>23.83</b>  |
|                    | bit2bit  | 0.1   | 17.89        | <b>44.77</b> | 24.11          | 17.73         | 14.28          | 17.06         |
|                    |          | 0.01  | 20.89        | 32.67        | <b>26.78</b>   | 24.43         | 15.67          | 19.29         |
|                    |          | 0.002 | 19.30        | 29.82        | 23.36          | 21.11         | 15.62          | 21.46         |
|                    | UWB      | 0.1   | 25.70        | 28.90        | 24.82          | 27.20         | 23.49          | 28.58         |
|                    |          | 0.01  | 20.58        | 28.83        | 22.42          | 16.74         | 14.18          | 21.13         |
|                    |          | 0.002 | 18.17        | 30.26        | 18.77          | 12.68         | 10.76          | 16.53         |
|                    | QNN      | 0.1   | 19.66        | 31.69        | 20.67          | 24.06         | 22.11          | 26.13         |
|                    |          | 0.01  | 19.32        | <b>37.14</b> | 24.95          | 13.09         | 9.88           | 15.01         |
|                    |          | 0.002 | 16.81        | <b>35.69</b> | 23.91          | 11.08         | 8.44           | 13.30         |
| LPIPS $\downarrow$ | proposed | 0.1   | <b>0.082</b> | <b>0.063</b> | <b>0.157</b>   | 0.217         | 0.450          | <b>0.083</b>  |
|                    |          | 0.01  | <b>0.151</b> | 0.080        | <b>0.263</b>   | 0.412         | 0.629          | <b>0.328</b>  |
|                    |          | 0.002 | <b>0.344</b> | 0.153        | 0.456          | 0.635         | 0.883          | <b>0.625</b>  |
|                    | bit2bit  | 0.1   | 0.328        | 0.139        | 0.287          | 0.248         | 0.397          | 0.302         |
|                    |          | 0.01  | 0.276        | 0.097        | 0.279          | <b>0.387</b>  | <b>0.479</b>   | 0.551         |
|                    |          | 0.002 | 0.440        | 0.138        | <b>0.434</b>   | <b>0.493</b>  | <b>0.624</b>   | 0.674         |
|                    | UWB      | 0.1   | 0.315        | 0.173        | 0.490          | 0.365         | 0.750          | 0.257         |
|                    |          | 0.01  | 0.570        | 0.278        | 0.855          | 0.708         | 1.088          | 0.533         |
|                    |          | 0.002 | 0.696        | 0.338        | 1.059          | 0.812         | 1.221          | 0.818         |
|                    | QNN      | 0.1   | 0.192        | 0.073        | 0.195          | <b>0.190</b>  | <b>0.340</b>   | 0.176         |
|                    |          | 0.01  | 0.271        | <b>0.056</b> | 0.331          | 0.478         | 0.677          | 0.471         |
|                    |          | 0.002 | 0.468        | <b>0.095</b> | 0.519          | 0.629         | 0.827          | 0.727         |
| SSIM $\uparrow$    | proposed | 0.1   | <b>0.913</b> | <b>0.909</b> | <b>0.812</b>   | <b>0.887</b>  | <b>0.782</b>   | <b>0.930</b>  |
|                    |          | 0.01  | <b>0.843</b> | 0.841        | <b>0.777</b>   | 0.781         | 0.653          | <b>0.746</b>  |
|                    |          | 0.002 | <b>0.624</b> | 0.493        | <b>0.625</b>   | 0.556         | 0.362          | <b>0.442</b>  |
|                    | bit2bit  | 0.1   | 0.545        | 0.900        | 0.441          | 0.759         | 0.670          | 0.545         |
|                    |          | 0.01  | 0.739        | 0.822        | 0.699          | <b>0.801</b>  | <b>0.716</b>   | 0.391         |
|                    |          | 0.002 | 0.541        | 0.558        | 0.513          | <b>0.688</b>  | <b>0.603</b>   | 0.350         |
|                    | UWB      | 0.1   | 0.740        | 0.810        | 0.561          | 0.745         | 0.458          | 0.822         |
|                    |          | 0.01  | 0.523        | 0.713        | 0.289          | 0.442         | 0.154          | 0.553         |
|                    |          | 0.002 | 0.402        | 0.664        | 0.113          | 0.351         | 0.084          | 0.321         |
|                    | QNN      | 0.1   | 0.749        | 0.884        | 0.436          | 0.859         | 0.767          | 0.828         |
|                    |          | 0.01  | 0.744        | <b>0.911</b> | 0.633          | 0.510         | 0.352          | 0.356         |
|                    |          | 0.002 | 0.573        | <b>0.843</b> | 0.539          | 0.284         | 0.168          | 0.186         |

## I.4 Velocity detection

This section evaluates the accuracy and robustness of our velocity detection method (Section 3.2 of the main paper and Section C) on controlled synthetic benchmarks where ground-truth motion is known exactly. We study four aspects of detector behavior:

1. **Best-detection accuracy:** how closely the top detected velocity matches the true velocity for a single moving Gaussian blob.
2. **Detection-set quality:** the precision of the detector.
3. **Velocity limits:** detector behavior as the true speed approaches the lower and upper bounds of the discretized search range.
4. **Multiple objects:** a multi-object extension with several independently moving blobs, comparing global scene-level detection against local object-centered detection.

Together, these experiments characterize how the detector responds to changes in photon rate, motion speed, and scene complexity.

### I.4.1 Accuracy of the best detection

**Scene and motion model:** We evaluate the accuracy of the best detected velocity on a synthetic benchmark consisting of a single translating Gaussian blob. The blob follows constant-velocity motion in the image plane, with velocity magnitude chosen from

$$\{0.5, 1.0, 1.5, 2.5\} \text{ px/frame},$$

and direction chosen from

$$\{0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ\}.$$

For each speed-direction pair, the blob trajectory is initialized so that the motion is approximately centered over the full sequence.

**Photon simulation and spatiotemporal volume:** For each motion configuration, we generate photon arrivals from the Gaussian flux model using the same synthetic event-generation pipeline as in the rest of the paper. The spatiotemporal volume has dimensions

$$256 \times 128 \times 128,$$

corresponding to time, height, and width, respectively. We repeat each configuration for 10 independent random seeds and for multiple photon levels by scaling the source amplitude. In the reported plots, the horizontal axis is the resulting photon rate, expressed in photons per second.

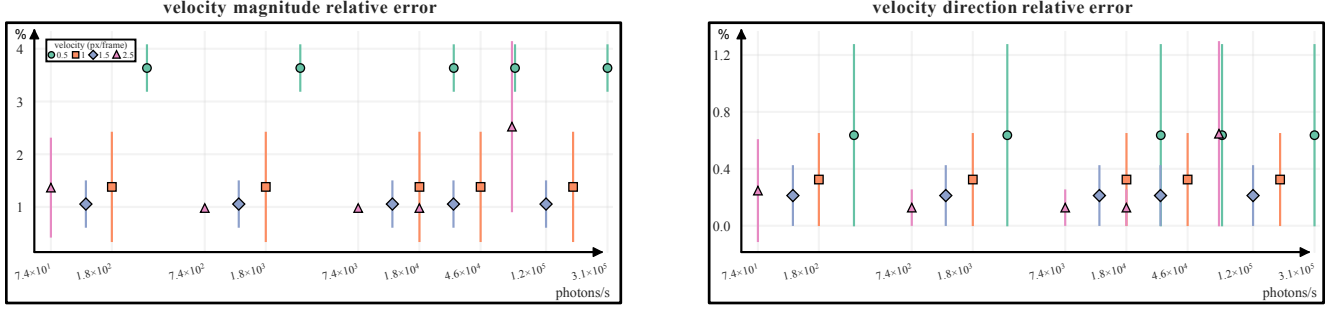
**Velocity search and detection:** For each simulated sequence, events are binned into a spatiotemporal tensor, transformed into the Fourier domain, and converted into a 2D velocity energy map using the plane-energy aggregation procedure described above. Candidate velocities are sampled uniformly over

$$[-3, 3] \times [-3, 3]$$

with 150 bins per axis. We then apply the local rank-based CFAR detector to this energy map with a window  $w = 100$  and a guard of  $g = 3$ . For each trial, we retain the detected velocity closest to the ground-truth motion and treat it as the best detection. Thus, this experiment isolates the intrinsic estimation accuracy of the detector once a plausible candidate has been produced, rather than emphasizing the total number of detections. In this sense, it measures the quality of the best-matched hypothesis, conditional on the presence of a reasonable candidate in the detection set.

**Reported quantities:** We report relative endpoint error as the magnitude metric and angular error as the directional metric. The total number of detections and inlier composition are analyzed separately in the following subsection.

**Velocity magnitude accuracy:** Figure I.6 shows that the best detected velocity is accurate in magnitude over a broad range of light levels. For intermediate velocities (1.0 and 1.5 px/frame), the mean relative endpoint error remains low and nearly flat, indicating that once a candidate is detected, its magnitude is recovered reliably. The 0.5 px/frame case is similarly stable but at a consistently higher error, while 2.5 px/frame produces the largest error across all photon rates. Even in that case, however, the error remains modest. Overall, the dominant source of variation is motion speed rather than photon rate: across the tested range, the curves are relatively flat, and the error level depends more strongly on the underlying velocity magnitude.



**Figure I.6: Accuracy of the best detected velocity for a single translating Gaussian blob.** **Left:** Mean relative endpoint error in velocity magnitude, expressed as a percentage of the ground-truth speed, versus light level. **Right:** Mean directional error, expressed as a percentage of  $180^\circ$ , versus light level. Markers correspond to different ground-truth velocities and error bars indicate one standard deviation over random seeds. The best detected velocity remains accurate and stable across the tested light level range, with the smallest errors at intermediate speeds (1.0–1.5 px/frame) and slightly larger errors at the slowest and fastest motions.

**Velocity directional accuracy:** The angular error in Figure I.6 follows the same trend. Intermediate speeds again yield the most accurate direction estimates, with small errors and weak photon-rate dependence. The 0.5 px/frame case is slightly worse but stable, while 2.5 px/frame produces the highest directional error, though the degradation is gradual. These results indicate that the method recovers both the magnitude and direction of the dominant motion with high fidelity once a correct candidate is present in the detection set.

**Interpretation:** Taken together, these results show that the main limitation of the method in this single-object setting is not the precision of the best detection once a reasonable candidate has been produced. Rather, the best-detection accuracy is already strong over a wide range of photon rates, especially for intermediate motion magnitudes. The most noticeable degradation occurs at the smallest and largest tested speeds. This is consistent with the intuition that very slow motions lie closer to the near-static regime, while larger motions are more affected by discretization and support effects in the sampled velocity domain. Even so, the best detected velocity remains close to the ground truth throughout the tested range.

## I.4.2 Detection set quality for a single moving object

**Scene and motion model:** We extend the previous single-object Gaussian-blob experiment to study not only the accuracy of the best detected velocity, but also the quality of the entire detection set. As before, the scene consists of a single translating Gaussian blob undergoing constant-velocity motion in the image plane. The tested speeds are

$$\{0.5, 1.0, 1.5, 2.5\} \text{ px/frame}, \quad (193)$$

and the tested directions are

$$\{0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ\}. \quad (194)$$

For each speed-direction pair, the trajectory is initialized so that the motion is approximately centered over the full sequence.

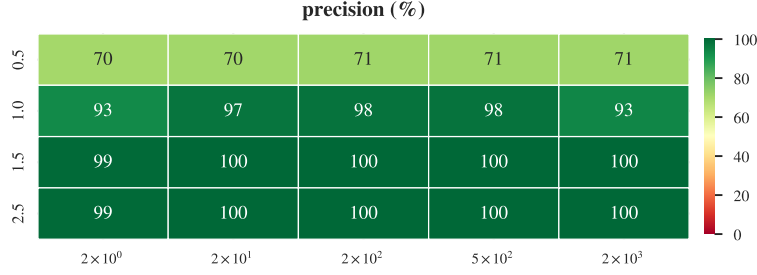
**Photon simulation and spatiotemporal volume:** For each motion configuration, we simulate photon arrivals from the same Gaussian flux model used in the previous experiment. The spatiotemporal volume has dimensions

$$256 \times 128 \times 128, \quad (195)$$

corresponding to time, height, and width, respectively. Each configuration is repeated for 10 independent random seeds and for multiple photon levels obtained by scaling the source amplitude. In the reported plots, the horizontal axis is the resulting photon rate, expressed in photons per second.

**Velocity search and detection:** For each simulated sequence, events are binned into a spatiotemporal tensor, transformed into the Fourier domain, and converted into a 2D velocity energy map using the plane-energy aggregation procedure described above. Candidate velocities are sampled uniformly over

$$[-3, 3] \times [-3, 3] \quad (196)$$



**Figure I.7: Precision of the velocity detection set**, defined as the fraction of detected hypotheses that are inliers, shown as a function of ground-truth speed (vertical axis, px/frame) and photon rate (horizontal axis, photons/s). Precision is near-perfect ( $\geq 93\%$ ) for speeds of 1.0 px/frame and above, and remains stable across photon rates. The 0.5 px/frame case shows lower precision ( $\sim 70\%$ ), consistent with the higher relative discretization error at slow speeds.

with 400 bins per axis. We then apply the local rank-based CFAR detector to this energy map with a window of 300 and a guard of 3. Unlike the previous subsection, where we retained only the closest detected velocity, here we keep the full set of detected velocity hypotheses for each trial.

**Reported quantities:** Let  $\mathcal{D}$  denote the full detection set for a given trial, and let  $\mathbf{v}_{\text{gt}}$  denote the ground-truth velocity. A detected hypothesis  $\hat{\mathbf{v}} \in \mathcal{D}$  is declared an inlier if it satisfies the hybrid criterion

$$\|\hat{\mathbf{v}} - \mathbf{v}_{\text{gt}}\|_2 \leq \max(\tau \|\mathbf{v}_{\text{gt}}\|_2, 3 \Delta v), \quad (197)$$

where  $\tau$  is the relative tolerance parameter and  $\Delta v$  is the spacing of the discrete velocity grid. In our implementation, we use  $\tau = 0.15$ . This hybrid tolerance is intended to avoid penalizing slow motions too harshly, since at small ground-truth speeds the grid quantization itself can represent a nonnegligible fraction of the motion magnitude.

Using this criterion, we compute three quantities for each trial:

1. the total number of detections,  $|\mathcal{D}|$ ;
2. the total number of inliers,  $|\mathcal{I}|$ ;
3. the total number of outliers,

$$|\mathcal{O}| = |\mathcal{D}| - |\mathcal{I}|. \quad (198)$$

**Detection precision:** Figure I.7 shows detection precision, defined as the percentage of detected hypotheses satisfying the inlier criterion. Precision depends primarily on speed rather than photon rate. At 0.5 px/frame, precision plateaus around 70–71% across all photon levels, indicating that a substantial fraction of detections fall outside the inlier neighborhood even though the total detection count is stable. This is attributable to two factors: slower objects produce broader Fourier-domain support that does not concentrate tightly on the ideal motion plane, and the fixed absolute grid spacing represents a larger fraction of the true speed. At 1.0 px/frame, precision rises to 93–98%, and for 1.5 and 2.5 px/frame it reaches 99–100%. Once the motion is sufficiently fast relative to the grid spacing, nearly all detections concentrate around the true velocity.

### I.4.3 Detection accuracy across the velocity range

**Goal of the experiment:** We next study the range of motion magnitudes over which the detector remains accurate under high illumination and single direction conditions. The goal is to isolate the effect of velocity itself, and in particular to probe failure modes near the lower and upper limits of the discretized search space.

**Scene and motion model:** We use the same synthetic single Gaussian blob benchmark as in the previous section, but fix the motion direction to  $0^\circ$ , so that the ground-truth velocity is always horizontal:

$$\mathbf{v}_{\text{gt}} = (v, 0), \quad v > 0.$$

This removes directional variation and allows the experiment to focus exclusively on the dependence of detection accuracy on speed.

**Ground-truth speed sweep:** We sample the ground-truth speed over the interval

$$v \in [0.5, 40] \text{ px/frame},$$

excluding  $v = 0$ . The sampling is chosen to be denser near the lower and upper ends of the range, where breakdown is most likely to occur, and coarser in the middle where the method is expected to be more stable. Concretely, we use dense sampling from 0.5 to 2 px/frame, a coarser sweep over intermediate speeds, and dense sampling again in the high-speed regime. This design is intended to resolve both the low-speed regime, where motion approaches the minimum nonzero scale represented by the search grid, and the high-speed regime, where large displacements and truncation effects may degrade performance.

**Photon simulation and spatiotemporal volume:** For all trials, we keep the light level fixed at the standard high flux setting used throughout the synthetic experiments. The spatiotemporal volume has dimensions

$$256 \times 128 \times 128,$$

and we use the same Gaussian blob model and event-generation pipeline as in the previous benchmark.

**Velocity search and detection:** The detector is run on the full 2D velocity search space, but the search domain is restricted to nonnegative velocities:

$$(v_x, v_y) \in [0, 40] \times [0, 40].$$

The grid spacing is fixed to

$$\Delta v = 0.25 \text{ px/frame},$$

which corresponds to uniform sampling of the first quadrant. Thus, while the ground-truth motion always lies on the horizontal axis, the detector is still free to return off-axis hypotheses, allowing the experiment to reveal both magnitude errors and deviations away from the true direction.

**Evaluation protocol:** Each speed is evaluated over 10 independent random seeds. For every trial, we run the standard plane-energy aggregation and rank-based CFAR detection pipeline, then identify the detected hypothesis nearest to the ground truth in the full 2D velocity plane.

**Reported quantities:** As the primary metric, we report the relative endpoint error

$$100 \cdot \frac{\|\hat{\mathbf{v}} - \mathbf{v}_{\text{gt}}\|_2}{\|\mathbf{v}_{\text{gt}}\|_2},$$

together with the directional error, expressed as a percentage of  $180^\circ$ . We also visualize the recovered horizontal component  $\hat{v}_x$  and the recovered vertical component  $\hat{v}_y$ . Since the true motion is purely horizontal, these two diagnostics make it possible to distinguish magnitude mismatch along the correct axis from off-axis leakage into the vertical component.

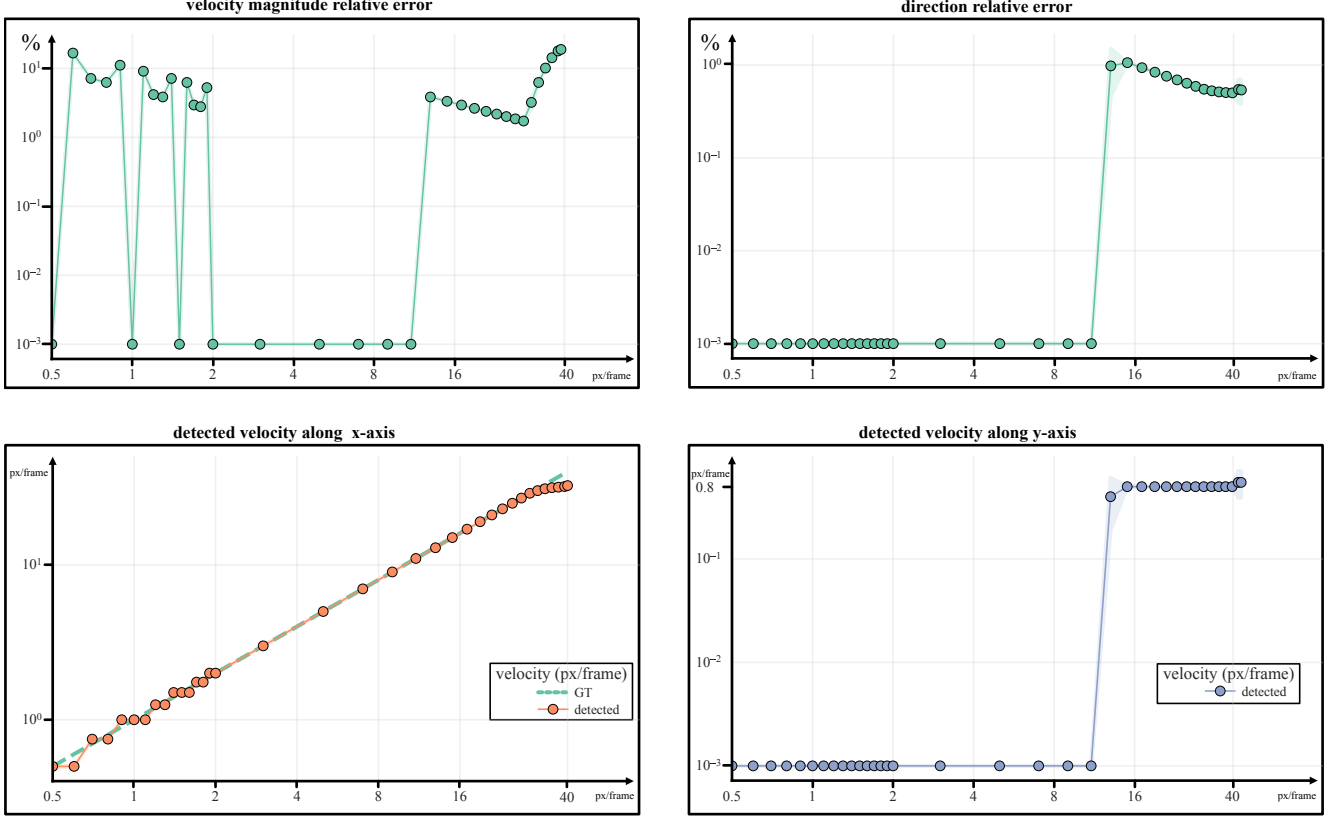
**Velocity detection across speed regimes:** Figure I.8 reveals three distinct regimes. In the intermediate range ( $\sim 2.5$ – $10$  px/frame), the detector is effectively exact: magnitude and directional errors fall to the numerical floor,  $\hat{v}_x$  follows the identity line, and  $\hat{v}_y$  remains zero. Below  $\sim 2$  px/frame, the relative error becomes irregular due to velocity-grid quantization—the absolute error remains small, but the relative metric is sensitive when  $\|\mathbf{v}_{\text{gt}}\|_2$  is comparable to the grid spacing.

**High-speed breakdown:** Above  $\sim 12$  px/frame, the detector undergoes a qualitative change. The recovered horizontal component continues to track the ground truth (Fig. I.8, bottom-left), but a persistent nonzero vertical component appears (Fig. I.8, bottom-right), producing a sharp jump in directional error while the magnitude error remains modest. This indicates that the dominant failure mode at high speed is not a loss of detectability but a systematic angular bias: the detector selects an off-axis hypothesis with nearly correct  $\hat{v}_x$  but spurious  $\hat{v}_y > 0$ . At the highest tested speeds, the magnitude error also increases, indicating that the breakdown extends beyond direction to affect the recovered speed. This transition likely reflects discretization or truncation effects in the Fourier-domain velocity grid that become significant for large per-frame displacements.

#### I.4.4 Detection accuracy with multiple moving objects

In this section, we evaluate the velocity detector on scenes containing multiple independently moving objects, testing whether it can recover several distinct velocity hypotheses from a single observation. We vary the number of objects from  $K = 2$  to  $K = 5$  and compare two detection strategies: global detection over the full scene and local detection within object-centered spatial crops.

**Scene and motion model:** We evaluate the detector in scenes containing multiple independently moving objects, rather than



**Figure I.8: Velocity-range benchmark for horizontal Gaussian-blob motion.** The ground-truth motion is purely horizontal with speed increasing from 0.5 to 40 px/frame. **Top left:** The relative magnitude error remains very small over an intermediate speed range, then rises at higher speeds as a spurious vertical component appears. **Top right:** The angular error follows the same transition and increases sharply once the estimate  $v_y > 0$ . **Bottom left:** The recovered horizontal component tracks the ground truth closely over most of the sweep and only deviates noticeably at the largest speeds. **Bottom right:** The recovered vertical component stays near zero at low and intermediate speeds, then exhibits a clear jump, indicating that the breakdown begins as an off-axis error before substantially affecting the magnitude and angle estimates themselves.

a single isolated blob. Each trial synthesizes a scene with

$$K \in \{2, 3, 4, 5\}$$

moving Gaussian blobs undergoing constant-velocity motion in the image plane. For object  $k \in \{1, \dots, K\}$ , the ground-truth image-plane velocity is denoted by  $\mathbf{v}_{\text{gt}}^{(k)} = (v_x^{(k)}, v_y^{(k)})$ . We sample  $\mathbf{v}_{\text{gt}}^{(k)}$  by drawing its speed uniformly from a prescribed interval  $[v_{\min}, v_{\max}]$  and its direction uniformly from  $[0, 360^\circ)$ , and then converting to Cartesian components. To avoid degenerate cases in which two objects have nearly identical motions, we enforce a minimum pairwise separation in velocity space:

$$\|\mathbf{v}_{\text{gt}}^{(k)} - \mathbf{v}_{\text{gt}}^{(\ell)}\|_2 \geq d_{\min} \quad \text{for all } k \neq \ell.$$

Objects are assigned unequal amplitudes using a fixed decreasing weight profile, so that each scene contains both stronger and weaker motion components. Spatially, each object is placed by sampling a midpoint location within the field of view and then back-projecting its start position from its velocity. This guarantees that each object passes through the scene during the observation window, while allowing objects to overlap spatially.

**Photon simulation and spatiotemporal volume:** The scene’s spatiotemporal flux is generated as a sum of  $K$  translating Gaussian blobs, each with its own velocity and amplitude. Photon arrivals are simulated from this composite spatiotemporal flux using the same synthetic event-generation pipeline as in the single-object experiments, and are then binned into a

spatiotemporal tensor of size

$$T \times H \times W = 256 \times 128 \times 128,$$

corresponding to time, height, and width, respectively. Each configuration ( $K$ , seed) is repeated over multiple random seeds, producing different realizations of velocities, layouts, and photon arrival randomness.

**Velocity search and detection:** For each simulated scene, events are binned into a spatiotemporal tensor, transformed into the Fourier domain, and converted into a 2D velocity energy map using the plane-energy aggregation procedure described above. Candidate velocities are sampled uniformly over

$$[-3, 3] \times [-3, 3]$$

with 150 bins per axis. We then apply the local rank-based CFAR detector to produce a binary detection set of velocity hypotheses.

We compare two motion-recovery strategies on the same simulated data.

- (i) **Global detection:** In the global setting, the detector is applied once to the full spatiotemporal volume of the scene, producing a single detection set shared by all objects. For each object, we assign the closest detected velocity in this global set and compute the corresponding recovery metrics.
- (ii) **Local detection:** In the local setting, we assume known object locations and extract a spatial crop centered on the midpoint of each object’s trajectory. The detection pipeline is then applied independently to each crop, yielding an object-specific detection set. For each object, we assign the nearest detected velocity from its local set and compute the same metrics as in the global case.

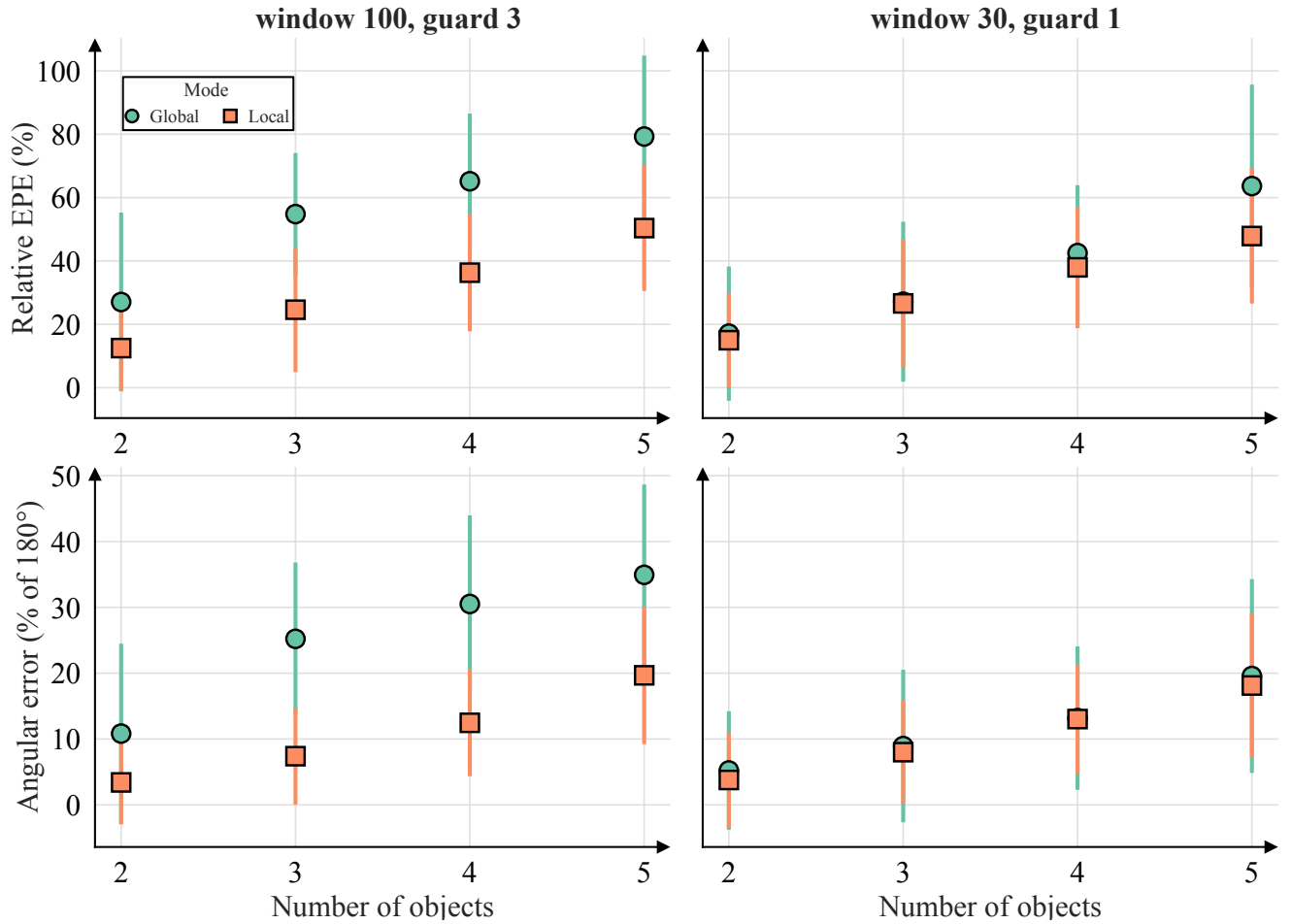
**Reported quantities:** For each object and each mode (global/local), we record the relative endpoint error and the angular error between the estimated and ground-truth velocities. Denoting the matched estimate by  $\hat{\mathbf{v}}^{(k)}$ , we report the relative endpoint error as a percentage of  $\|\mathbf{v}_{\text{gt}}^{(k)}\|_2$ , and the angular error as a percentage of  $180^\circ$ . A recovery is declared successful when the matched estimate has relative endpoint error below 10%. For the plots in Fig. 1.9, results are first averaged over objects within each scene and then averaged over random seeds, yielding scene-level summaries as functions of  $K$ .

**Detector neighborhood sweep:** To assess the effect of CFAR support size in multi-motion scenes, we repeat the experiment for two detector neighborhood settings: a larger neighborhood with window size 100 and guard size 3, and a smaller neighborhood with window size 30 and guard size 1. The two columns of Fig. 1.9 compare these settings directly.

**Results:** Figure 1.9 shows that global detection degrades steadily as scene complexity increases: both relative endpoint error and angular error grow with  $K$ , reaching high values at  $K = 5$ . This is expected, since the global detector aggregates evidence from all objects simultaneously, and competing motion signatures interfere in the velocity domain.

Local detection consistently outperforms global detection across all settings, confirming that spatial localization suppresses cross-object interference. The benefit of localization is most pronounced with the larger CFAR neighborhood (window 100), where the global method aggregates over a broader context and is therefore more susceptible to contamination. With the smaller neighborhood (window 30), the gap narrows, indicating that a more localized CFAR support already mitigates some of the interference even without explicit spatial cropping.

These results motivate the practical pipeline demonstrated in our real experiments: applying velocity detection to spatially localized regions of the scene, where each region is dominated by a single dominant motion.



**Figure I.9: Velocity detection with multiple moving objects.** Relative endpoint error (top) and angular error (bottom) as a function of the number of objects  $K$ , for global detection (green circles) and local detection on object-centered crops (orange squares). Left and right columns compare two CFAR neighborhood sizes (window 100, guard 3 vs. window 30, guard 1). Error bars indicate one standard deviation over random seeds. Local detection consistently outperforms global detection, with the gap widening as  $K$  increases. A smaller CFAR neighborhood partially closes the gap by reducing cross-object interference in the global setting.

## References

- [1] J.B. Allen and L.R. Rabiner. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE*, 65(11):1558–1564, 1977. 33
- [2] Andrew C. Berry. The Accuracy of the Gaussian Approximation to the Sum of Independent Variates. *Trans. Amer. Math. Soc.*, 49(1):122–136, 1941. 42
- [3] Prateek Chennuri, Yiheng Chi, Enze Jiang, G. M. Dilshan Godaliyadda, Abhiram Gnanasambandam, Hamid R. Sheikh, Istvan Gyongy, and Stanley H. Chan. Quanta Video Restoration. In *Proc. Eur. Conf. on Computer Vision (ECCV)*, pages 152–171. Springer, 2024. 44
- [4] Samuel N. Cohen and Robert J. Elliott. *Stochastic Calculus and Applications*. Birkhäuser New York, 2nd edition, 2015. 20
- [5] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, 19(90):297–301, 1965. 33
- [6] R. Crochiere. A weighted overlap-add method of short-time Fourier analysis/Synthesis. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(1):99–102, 1980. 33, 34
- [7] Daryl J. Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2003. 4
- [8] Dawei W. Dong and Joseph J. Atick. Statistics of natural time-varying images. *Netw. Comput. Neural Syst.*, 6(3):345–358, 1995. 11, 12
- [9] Carl-Gustav Esseen. *On the Liapounoff Limit of Error in the Theory of Probability*. Almqvist & Wiksell, 1942. 42
- [10] Yang Feng, Hengyi Lv, Hailong Liu, Yisa Zhang, Yuyao Xiao, and Chengshan Han. Event Density Based Denoising Method for Dynamic Vision Sensor. *Appl. Sci.*, 10(6), 2020. 26
- [11] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 4th edition, 2018. 33
- [12] Joseph W. Goodman. *Statistical Optics*. John Wiley & Sons, 1985. 4
- [13] F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*, 66(1):51–83, 1978. 33, 34
- [14] 8K VIDEOS ULTRA HD. Super SLOW MOTION in 8K ULTRA HD 1000FPS. YouTube Video, <https://www.youtube.com/watch?v=LrzWrvOjJm8>, 2020. Accessed: 2026-06-30. 44
- [15] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From Video Frames to Realistic DVS Events. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1312–1321, 2021. 26
- [16] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, Jianing Li, Shanshan Jia, Yihua Fu, Boxin Shi, Si Wu, and Yonghong Tian. 1000× Faster Camera and Machine Vision with Ordinary Devices. *Engineering*, 25:110–119, 2023. 20, 36
- [17] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-Time Intermediate Flow Estimation for Video Frame Interpolation. In *Proc. Eur. Conf. on Computer Vision (ECCV)*, pages 624–642. Springer, 2022. 44
- [18] iniVation. DAVIS 240 Datasheet. <https://inivation.com/wp-content/uploads/2019/08/DAVIS240.pdf>, 2019. Accessed: 2026-06-30. 37
- [19] Monique Jeanblanc, Marc Yor, and Marc Chesney. *Mathematical methods for financial markets*. Springer Science & Business Media, 2009. 4, 8
- [20] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.*, 82(1):35–45, 1960. 36, 37
- [21] Steven M. Kay. Statistical signal processing: estimation theory. *Prentice Hall*, 1993. 16
- [22] Alireza Khodamoradi and Ryan Kastner. O(N)-Space Spatiotemporal Filter for Reducing Noise in Neuromorphic Vision Sensors. *IEEE Trans. Emerg. Topics Comput.*, 9(1):15–23, 2021. 26
- [23] John F. C. Kingman. *Poisson Processes*. Oxford University Press, 1993. 9
- [24] Günter Last and Mathew Penrose. *Lectures on the Poisson Process*. Cambridge University Press, 2017. 6, 9, 10, 27
- [25] Erich Leo Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer, 2005. 18
- [26] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A  $128 \times 128$  120 dB 15  $\mu$ s latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008. 26
- [27] Yehe Liu. SPADTools: SPAD512 Data Loader and Preprocessor. <https://github.com/lyehe/spadtools>, 2024. GitHub repository. Accessed: 2026-06-30. 38
- [28] Yehe Liu, Alexander Krull, Hector Basevi, Aleš Leonardis, and Michael Jenkins. bit2bit: 1-bit quanta video reconstruction via self-supervised photon prediction. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 88443–88485, 2024. 30, 34, 38, 44
- [29] Sizhuo Ma, Shantanu Gupta, Arin Can Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Trans. Graph.*, 39(4):79:1–79:16, 2020. 40, 44
- [30] Xiangwei Meng. Rank sum nonparametric CFAR detector in nonhomogeneous background. *IEEE Trans. Aerosp. Electron. Syst. (TAES)*, 57(1):397–403, 2021. 17
- [31] Micro Photon Devices. PDM Series: Single Photon Avalanche Diode Modules. <https://www.micro-photon-devices.com/products/single-pixels/pdm>, 2025. Accessed: 2026-06-30. 33
- [32] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. *Int. J. Robot. Res.*, 36(2):142–149, 2017. 37, 40
- [33] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991. 9
- [34] Pi Imaging. SPAD 512. <https://piimaging.com/spad-512/>, 2021. Accessed: 2026-06-25. 33

- [35] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time Intensity Estimation Using Event Cameras. In *Proc. Asian Conf. on Computer Vision (ACCV)*, pages 308–324. Springer, 2018. [23](#), [24](#), [36](#)
- [36] Donald L. Snyder and Michael I. Miller. *Random point processes in time and space*. Springer Science & Business Media, 2012. [38](#)
- [37] Varun Sundar, Andrei Ardelean, Tristan Swedish, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Sodacam: Software-defined cameras via single-photon imaging. In *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 8165–8176, 2023. [36](#)
- [38] Varun Sundar, Matthew Dutson, Andrei Ardelean, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Generalized event cameras. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 25007–25017, 2024. [43](#)
- [39] Varun Sundar, Tianyi Zhang, Sacha Jungerman, and Mohit Gupta. Quanta neural networks: From photons to perception. In *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 5091–5101, 2025. [44](#)
- [40] David Vernon. *Fourier Vision: Segmentation and Velocity Measurement using the Fourier Transform*. Springer, 2001. [36](#)
- [41] Mian Wei, Sotiris Nousias, Rahul Gulve, David B Lindell, and Kiriakos N Kutulakos. Passive ultra-wideband single-photon imaging. In *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 8135–8146, 2023. [6](#), [8](#), [31](#), [34](#), [39](#), [44](#)
- [42] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like Visual Image Reconstruction via Spiking Neural Model. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1435–1443, 2020. [40](#)