Adapting Lightweight Image-based Counting Models for Video Crowd Counting

Supplementary Material

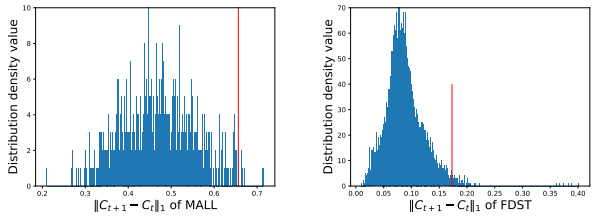

Figure 1. Examples of distributions of $\|C_{t+1} - C_t\|_1$. Red line is the 95% quantile.

A. Robust Estimation of the Temporal Variation Bound ϵ^* (Sec. 3.1)

The definition of ϵ^* in Eq. (3) of the main paper follows the classical principle of estimating a population-level upper bound from finite samples. However, directly adopting the empirical maximum over adjacent-frame L1 ChF distances may lead to instability: the maximum is often dominated by rare outlier frames caused by annotation noise, sudden camera jitter, or atypical scene transitions. Such artifacts do not represent meaningful temporal dynamics of crowd flow and therefore would distort the intended statistical prior.

To obtain a robust and distribution-consistent estimate of the intrinsic temporal variation, we replace the empirical maximum with a high quantile of the adjacent-frame distance distribution. Fig. 1 visualizes the distribution on MALL [1] and FDST [3]. Instead of using the 100% quantile, we adopt the 95% quantile (marked in red) as the default value of ϵ^* in all experiments. This choice retains the vast majority of statistically plausible temporal variations while effectively excluding noisy extremes.

This strategy is theoretically aligned with our formulation:

- the bound ϵ^* is meant to reflect statistically admissible variation,
- not the worst-case fluctuation caused by anomalies,
- and therefore should be estimated at the population scale, not dictated by a single outlier sample.

To empirically verify the stability of this robust estimator, we conduct an ablation study using different quantile levels. Results in Table 1 show that:

1. Performance is stable across a wide range of high quantiles (90–97.5%), indicating that the ST prior is not sensitive to small changes in ϵ^*;
2. The 95% quantile yields the best performance, demonstrating an effective balance between capturing true temporal dynamics and suppressing noise-driven deviations;
3. The full 100% maximum substantially degrades performance, confirming its susceptibility to outliers.

ϵ^*	90%		92.5%		95%		97.5%		100%	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
	14.7	21.4	14.3	20.6	**14.1**	**19.9**	14.4	19.9	16.1	24.8

Table 1. Ablation on the choice of quantile on DRONECROWD [8]. We evaluate different quantile levels of the adjacent-frame L1 ChF distance distribution when estimating ϵ^*.

Together, these findings validate our use of a quantile-based estimator for ϵ^* as a principled, data-driven, and robust approximation to the intrinsic temporal variation bound.

B. Visual Results

To complement the quantitative results reported in the paper, we provide qualitative density-map visualizations on VENICE [5] and DRONECROWD [8]. These two benchmarks represent two of the most challenging conditions in our evaluation: VENICE has very limited training data and relatively low frame rates, while DRONECROWD features aerial viewpoints where people appear at extremely small scales and the test scenes differ substantially from the training scenes. Together, they cover complementary difficulty regimes—data scarcity and severe visual scale reduction—and thus serve as representative examples for examining model behavior under challenging temporal dynamics. Fig. 2 and Fig. 3 show consecutive frames with ground-truth density maps and predictions from models trained with and without our statistical regularizer (SR).

Across both datasets, three consistent observations emerge.

1. With the proposed SR, the predictions become closer to the ground truth at both global scale and in the two cropped local regions. This behavior aligns with our formulation: by penalizing statistically implausible temporal deviations, the regularizer constrains the model to search within solutions of controlled complexity, which improves video crowd counting (VCC) accuracy without architectural changes.
2. In both figures, the predictions obtained without the SR exhibit larger frame-to-frame fluctuations in total counts, both at the global and local level. Since the count variation in the ground truth is comparatively small, these fluctuations are unlikely to stem solely from true scene

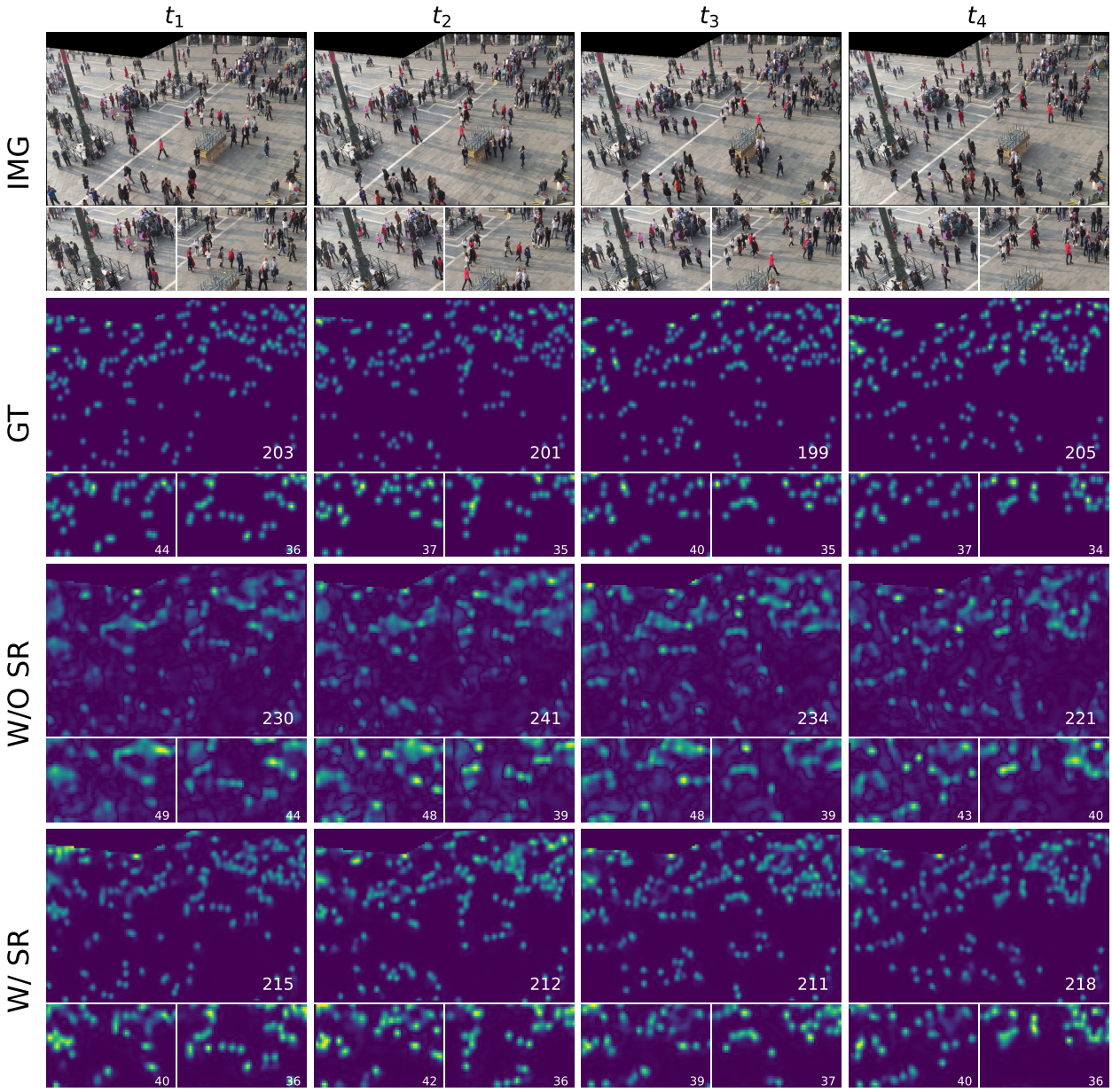

Figure 2. Density map visualization on the VENICE [5] dataset. Each column corresponds to a continuous time point t_1, t_2, t_3, t_4. For each time point, we display a 3-image composite in a pyramid layout: the full frame is shown on top, and two spatially cropped local regions from the same frame are placed at the bottom-left and bottom-right. The first row ("IMG") shows the raw images; the second row ("GT") shows the corresponding ground-truth density maps; the third row ("W/O SR") shows the predictions generated by a model trained without our statistical regularizer; and the fourth row ("W/ SR") shows the predictions of a model trained with our regularizer. For all density maps, the total predicted count is annotated in the lower-right corner.

motion. The reduced variation observed with the proposed SR aligns with our formulation: predictions violating the data-driven temporal bound are penalized during training. This encourages temporal behavior consistent with the intrinsic statistics of the dataset.

3. Local density maps further indicate that the SR improves the temporal coherence of spatial density patterns. In both datasets, predictions without SR sometimes show unstable local responses—such as regions becoming noticeably fainter, less distinct, or briefly disappearing in

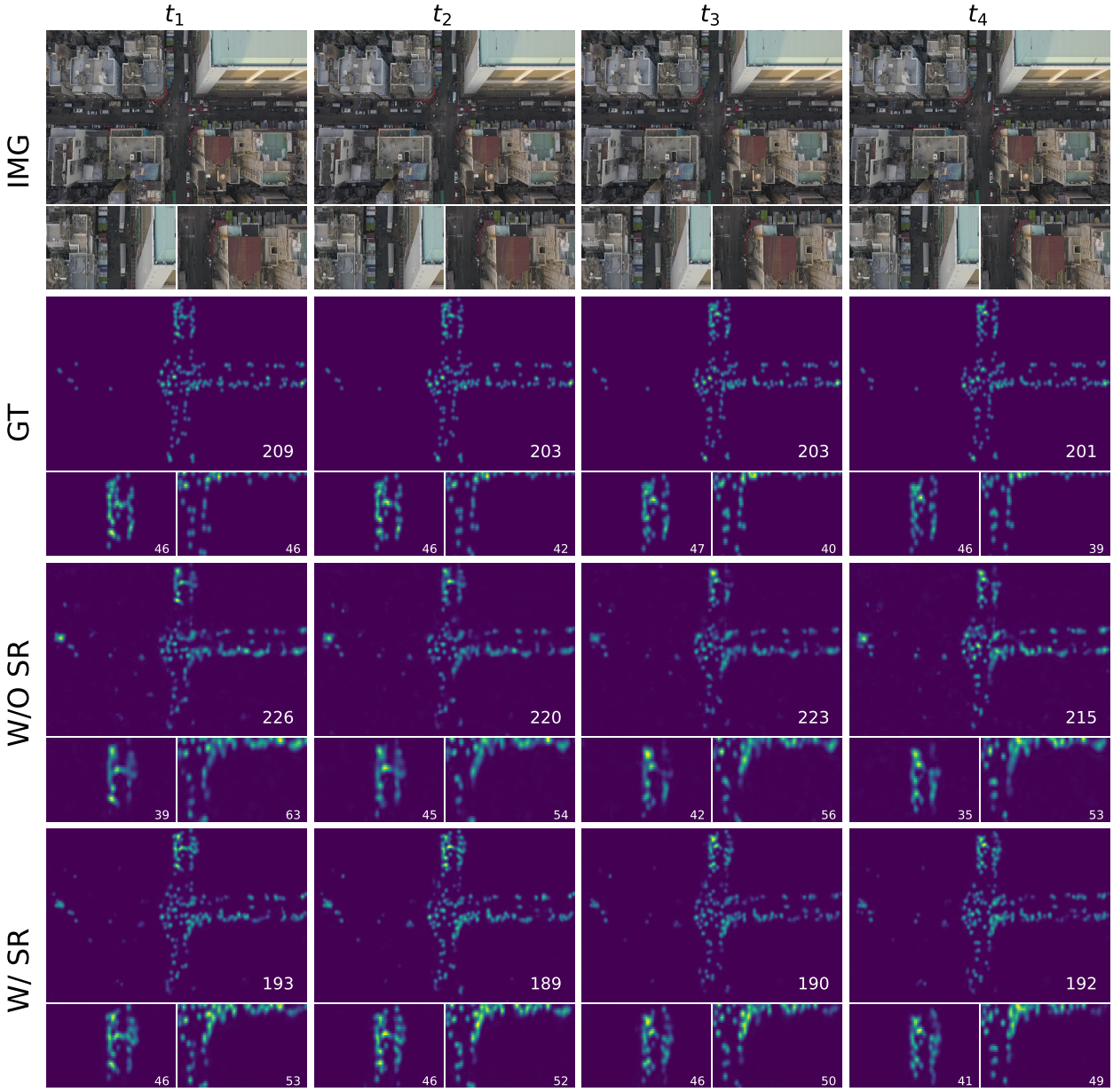

Figure 3. Density map visualization on the DRONECROWD [8] dataset. Each column corresponds to a continuous time point t_1, t_2, t_3, t_4. For every time point, we construct a 3-image composite showing the full frame (top) along with two spatially cropped local regions (bottom-left and bottom-right). The first row ("IMG") presents the raw images; the second row ("GT") shows the ground-truth density maps; the third row ("W/O SR") shows predictions from the model trained without our statistical regularizer; and the fourth row ("W/ SR") shows predictions from the model trained with our regularizer. Total people counts are displayed at the lower-right corner of each density map.

later frames despite being present in the ground truth. In contrast, the SR leads to smoother and more consistent local density evolution that better reflects the underlying scene dynamics. These observations demonstrate that the SR enhances temporal behavior in local spatial structure without introducing any inference overhead.

C. Proofs

C.1. Proof of Theorem 1

According to the definition in [6], the Characteristic function (ChF) value at $\mathbf{w}$ in the frequency domain at time t is

$$C_t(\mathbf{w}) = \sum_{i=1}^{Q} \exp(\mathrm{i}\langle \mathbf{x}_t^{(i)}, \mathbf{w}\rangle - \frac{\gamma^2}{2}\langle \mathbf{w}, \mathbf{w}\rangle) \quad (1)$$

where i is the imaginary unit, $\langle\cdot,\cdot\rangle$ means the inner product between two vectors, $\mathbf{x}_t^{(i)}$ is ith person's spatial position at time t, and γ is the bandwidth of the density map, if $\gamma = 0$, then it means the ChF of the dot map. Note that $\mathbf{w}$ is a coordinate at the frequency domain, i.e., a vector in the 2D Euclidean space. Therefore when it appears in $\langle\cdot,\mathbf{w}\rangle$, $\mathbf{w}^T$, and $\|\mathbf{w}\|_2$, we regard it as a 2D vector. Since the function $C_t(\mathbf{w})$ is smooth regarding $(\mathbf{x}_t^{(1)}, \cdots, \mathbf{x}_t^{(Q)})^T$, we take its derivative w.r.t. $\mathbf{X}_t = (\mathbf{x}_t^{(1)}, \cdots, \mathbf{x}_t^{(Q)})^T$ to get

$$\begin{aligned}\frac{\partial C_t(\mathbf{w})}{\partial \mathbf{X}_t} &= (\exp(\mathrm{i}\langle \mathbf{x}_t^{(1)}, \mathbf{w}\rangle - \frac{\gamma^2}{2}\langle \mathbf{w}, \mathbf{w}\rangle)\cdot \mathrm{i}\cdot \mathbf{w}^T, \ldots, \\ &\quad \exp(\mathrm{i}\langle \mathbf{x}_t^{(Q)}, \mathbf{w}\rangle - \frac{\gamma^2}{2}\langle \mathbf{w}, \mathbf{w}\rangle)\cdot \mathrm{i}\cdot \mathbf{w}^T) \\ &= A_{t,\mathbf{w}}\ . \end{aligned} \quad (2)$$

Notice that

$$\begin{aligned}\|A_{t,\mathbf{w}}\|_2 &= |\exp(-\frac{\gamma^2}{2}\langle \mathbf{w}, \mathbf{w}\rangle)\cdot \mathrm{i}|\cdot \|(\exp(\mathrm{i}\langle \mathbf{x}_t^{(1)}, \mathbf{w}\rangle)\cdot \mathbf{w}^T, \\ &\quad \ldots, \exp(\mathrm{i}\langle \mathbf{x}_t^{(Q)}, \mathbf{w}\rangle)\cdot \mathbf{w}^T)\|_2 \quad (3)\\ &\le \|(\exp(\mathrm{i}\langle \mathbf{x}_t^{(1)}, \mathbf{w}\rangle)\cdot \mathbf{w}^T, \ldots, \\ &\quad \exp(\mathrm{i}\langle \mathbf{x}_t^{(Q)}, \mathbf{w}\rangle)\cdot \mathbf{w}^T)\|_2 \quad (4)\\ &\le \sum_{i=1}^{Q} \|\exp(\mathrm{i}\langle \mathbf{x}_t^{(i)}, \mathbf{w}\rangle)\cdot \mathbf{w}^T\|_2 \quad (5)\\ &\le \sum_{i=1}^{Q} |\exp(\mathrm{i}\langle \mathbf{x}_t^{(i)}, \mathbf{w}\rangle)|\cdot \|\mathbf{w}\|_2 \quad (6)\\ &\le \sum_{i=1}^{Q} \|\mathbf{w}\|_2 \le Q\|\mathbf{w}\|_2. \quad (7)\end{aligned}$$

(3), (6) hold by absolute homogeneity of the norm, (4) is due to $|\exp(-\frac{\gamma^2}{2}\langle \mathbf{w}, \mathbf{w}\rangle)\cdot \mathrm{i}| \le 1$, (5) is the direct calculation, (7) is because $|\exp(\mathrm{i}\langle \mathbf{x}_t^{(i)}, \mathbf{w}\rangle)| = 1$.

Now by the smoothness of $C_t(\mathbf{w})$ w.r.t. $\mathbf{X}_t$, we have the following convergence when $\Delta t \to 0$:

$$\frac{|\ C_{t+\Delta t}(\mathbf{w}) - C_t(\mathbf{w}) - A_{t,\mathbf{w}}(\mathbf{X}_{t+\Delta t} - \mathbf{X}_t)\ |}{|\ \mathbf{X}_{t+\Delta t} - \mathbf{X}_t\ |} \to 0. \quad (8)$$

Since the denominator tends to zero as $\Delta t \to 0$, the numerator also tends to zero. Therefore, rearranging the term in numerator, we have

$$C_{t+\Delta t}(\mathbf{w}) - C_t(\mathbf{w}) \xrightarrow{\Delta t \to 0} A_{t,\mathbf{w}}(\mathbf{X}_{t+\Delta t} - \mathbf{X}_t) \quad (9)$$

where $\|A_{t,\mathbf{w}}\|_2 \le Q\|\mathbf{w}\|_2$ as we proved in (7).

C.2. Proof of Theorem 2

This theorem provides an upper bound on the local change of the density map by leveraging the temporal evolution of the characteristic function between consecutive frames. While Proposition 2 in [6] offers a static upper bound on the spatial difference between density maps, this result extends the idea to a temporal context.

From Proposition 2 in [6], we know that:

$$\int_R D_{t+1}(\mathbf{x}) - D_t(\mathbf{x})\, d\mathbf{x} \le (2\pi)^{-2}\|C_{t+1} - C_t\|_1 m(R). \quad (10)$$

Given the assumption that $\|C_{t+1} - C_t\|_1 \le \epsilon$, we can now write:

$$\int_R D_{t+1}(\mathbf{x}) - D_t(\mathbf{x})\, d\mathbf{x} \le (2\pi)^{-2}\epsilon m(R). \quad (11)$$

Now, dividing both sides by $m(R)$, we obtain:

$$\Delta_R(t, t+1) = \frac{\left|\int_R D_{t+1}(\mathbf{x}) - D_t(\mathbf{x})\, d\mathbf{x}\right|}{m(R)} \le (2\pi)^{-2}\epsilon. \quad (12)$$

Thus, we conclude that the local change in people count between consecutive frames is bounded by $(2\pi)^{-2}\epsilon$, which is directly related to the temporal variation of the characteristic function.

C.3. Proof of Theorem 3

Let $\Delta(\mathbf{w}) = C_{t+1}(\mathbf{w}) - C_t(\mathbf{w})$, and suppose the position of each individual j in I_t is $\mathbf{p}_j$, the movement of individual j from I_t to I_{t+1} is $\mathbf{m}_j$. We define the following symbol:

$$\delta_{\mathbf{p}_j,\mathbf{m}_j}(\mathbf{w}) = \exp(\mathrm{i}(\mathbf{p}_j + \mathbf{m}_j)^T\mathbf{w} - \tfrac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}) \quad (13)$$

$$\delta_{\mathbf{p}_j}(\mathbf{w}) = \exp(\mathrm{i}(\mathbf{p}_j)^T\mathbf{w} - \tfrac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}) \quad (14)$$

$$\mathbf{m}_j \sim \mathcal{L}_\infty \quad (15)$$

where i is the imaginary unit and $\mathcal{L}_\infty$ represents the asymptotic distribution of people motion. Then according to the definition of the characteristic function in [6], we have

$$\begin{aligned}&\mathrm{std}_{\mathbf{m}_j\sim\mathcal{L}_\infty}[\Delta(\mathbf{w})] \\ &= \sqrt{\mathrm{var}\left(\sum_{j=1}^{Q}\delta_{\mathbf{p}_j,\mathbf{m}_j}(\mathbf{w}) - \sum_{j=1}^{Q}\delta_{\mathbf{p}_j}(\mathbf{w})\right)} \quad (16)\\ &= \sqrt{\mathrm{var}\left(\sum_{j=1}^{Q}\delta_{\mathbf{p}_j,\mathbf{m}_j}(\mathbf{w})\right)} \quad (17)\end{aligned}$$

$$= \sqrt{\mathrm{var}\Big(\sum_{j=1}^{Q} \exp(\mathrm{i}\mathbf{p}_j^T\mathbf{w}) \exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w}) \exp(\frac{-\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}{2})\Big)} \tag{18}$$

$$= \sqrt{\sum_{j=1}^{Q} \big|\exp(\mathrm{i}\mathbf{p}_j^T\mathbf{w}) \exp(\frac{-\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}{2})\big|^2 \mathrm{var}(\exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w}))} \tag{19}$$

$$= \sqrt{\sum_{j=1}^{Q} \exp(-\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}) \mathrm{var}(\exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w}))}. \tag{20}$$

(17) holds because $\delta_{\mathbf{p}_j}(\mathbf{w}) = \sum_{j=1}^{M} \exp(\mathrm{i}\mathbf{p}_j^T\mathbf{w} - \frac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w})$ is a constant w.r.t. $\mathbf{m}_j$ in the variance. (19) holds by the fact that $\mathrm{var}(\sum_j c_j r_j) = \sum_j |c_j|^2 \mathrm{var}(r_j)$ for complex coefficients c_j and independent complex r.v.s r_j, and (20) holds by $|\exp(\mathrm{i}\mathbf{p}_j^T\mathbf{w})|^2 = 1$. For the variance term $\mathrm{var}_{\mathbf{m}_j \sim \mathcal{L}_\infty}(\exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w}))$ in (20), we need a series of lemmas to reveal the asymptotic distribution $\mathcal{L}_\infty$.

Lemma 1 *Suppose $f : \mathbb{R}^n \to \mathbb{C}$ is an integrable function, then the absolute integral inequality holds for complex-valued functions, i.e., $|\int_A f(s)\, ds| \le \int_A |f(s)|\, ds$, $\forall A \subset \mathbb{R}^n$ where n is any positive integer and A is any Lebesgue measurable set.*

Proof of Lemma 1 : Without loss of generality, we can assume $f(s) \neq 0$ *on* A, as $\int_A f(s)\, ds = \int_{A\cap\{f(s)=0\}} f(s)\, ds + \int_{A\cap\{f(s)\neq 0\}} f(s)\, ds = \int_{A\cap\{f(s)\neq 0\}} f(s)\, ds$ and similarly $\int_A |f(s)|\, ds = \int_{A\cap\{f(s)\neq 0\}} |f(s)|\, ds$. Let $\Re f$ and $\Im f$ represents the real and imaginary part of f respectively. By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\int_A |\Re f(s)|\, ds &= \int_A \frac{|\Re f(s)|}{|f(s)|^{\frac{1}{2}}} |f(s)|^{\frac{1}{2}}\, ds \\ &\le \sqrt{\int_A \frac{|\Re f(s)|^2}{|f(s)|}\, ds} \sqrt{\int_A |f(s)|\, ds}. \end{aligned} \tag{21}$$

Similarly,

$$\int_A |\Im f(s)|\, ds \le \sqrt{\int_A \frac{|\Im f(s)|^2}{|f(s)|}\, ds} \sqrt{\int_A |f(s)|\, ds}. \tag{22}$$

Then we have

$$\begin{aligned}&\Big|\int_A f(s)\, ds\Big|^2 \\ &= \Big|\int_A \Re f(s)\, ds\Big|^2 + \Big|\int_A \Im f(s)\, ds\Big|^2 \\ &\le \Big(\int_A |\Re f(s)|\, ds\Big)^2 + \Big(\int_A |\Im f(s)|\, ds\Big)^2 \\ &\le \int_A \frac{|\Re f(s)|^2}{|f(s)|}\, ds \int_A |f(s)|\, ds \\ &\quad + \int_A \frac{|\Im f(s)|^2}{|f(s)|}\, ds \int_A |f(s)|\, ds \\ &\le \int_A |f(s)|\, ds \left(\int_A \frac{|\Re f(s)|^2 + |\Im f(s)|^2}{|f(s)|}\, ds\right) \\ &\le \int_A |f(s)|\, ds \left(\int_A \frac{|f(s)|^2}{|f(s)|}\, ds\right) \\ &\le \Big(\int_A |f(s)|\, ds\Big)^2. \end{aligned} \tag{23}$$

Take the square root then we have the desired inequality. ■

Lemma 2 *Suppose X is a random variable taking values on $\mathbb{R}^k$ where k is some positive integer, then for $\forall \mathbf{t} \in \mathbb{R}^k$ and any positive integer n we have $|\mathbb{E}[\exp(\mathrm{i}\langle \mathbf{t}, X\rangle)] - \mathbb{E}[\sum_{m=0}^{n} \frac{(\mathrm{i}\langle \mathbf{t}, X\rangle)^m}{m!}]| \le \mathbb{E}[\min(|\langle \mathbf{t}, X\rangle|^{n+1}, 2|\langle \mathbf{t}, X\rangle|^n)]$.*

Proof of Lemma 2 : Fix $x \in \mathbb{R}$, we have the following Riemann integral equation by integration by parts:

$$\begin{aligned}&\int_0^x (x-s)^n \exp(\mathrm{i}s)\, ds \\ &= \int_0^x (x-s)^n \cos s\, ds + \mathrm{i}\int_0^x (x-s)^n \sin s\, ds \\ &= \frac{-1}{n+1}\Big[(x-s)^{n+1}\cos s\Big|_0^x + \int_0^x (x-s)^{n+1}\sin s\, ds \\ &\quad + \mathrm{i}(x-s)^{n+1}\sin s\Big|_0^x - \mathrm{i}\int_0^x (x-s)^{n+1}\cos s\, ds\Big] \\ &= \frac{-1}{n+1}\left[-x^{n+1} - \mathrm{i}\int_0^x (x-s)^{n+1}(\cos s + \mathrm{i}\sin s) ds\right] \\ &= \frac{x^{n+1}}{n+1} + \frac{\mathrm{i}}{n+1}\int_0^x (x-s)^{n+1}\exp(\mathrm{i}s)\, ds. \end{aligned} \tag{24}$$

Now taking $n = 0$ at first and then iteratively using (24) we have

$$\begin{aligned}&\int_0^x \exp(\mathrm{i}s)\, ds \\ &= x + \mathrm{i}\int_0^x (x-s)\exp(\mathrm{i}s)\, ds \\ &= x + \frac{\mathrm{i}x^2}{2} + \frac{\mathrm{i}^2}{2}\int_0^x (x-s)^2\exp(\mathrm{i}s)\, ds \\ &= \cdots \\ &= \sum_{m=1}^{n} \frac{\mathrm{i}^{m-1}x^m}{m!} + \frac{\mathrm{i}^n}{n!}\int_0^x (x-s)^n \exp(\mathrm{i}s)\, ds. \end{aligned} \tag{25}$$

Since $\int_0^x \exp(\mathrm{i}s) ds = \frac{\exp(\mathrm{i}x)-1}{i}$, together with (25) we

have

$$\exp(\mathrm{i}x) - \sum_{m=0}^{n} \frac{(\mathrm{i}x)^m}{m!} = \frac{\mathrm{i}^{n+1}}{n!} \int_0^x (x-s)^n \exp(\mathrm{i}s)\, ds. \tag{26}$$

By Lemma 1, we have

$$\begin{aligned}
&\left| \frac{\mathrm{i}^{n+1}}{n!} \int_0^x (x-s)^n \exp(\mathrm{i}s)\, ds \right| \\
&= \left| \frac{\mathrm{i}^{n+1}}{n!} \right| \left| \int_0^x (x-s)^n \exp(\mathrm{i}s)\, ds \right| \\
&\le \frac{1}{n!} \int_0^x |(x-s)^n \exp(\mathrm{i}s)|\, ds \\
&\le \frac{1}{n!} \left| \int_0^x (x-s)^n\, ds \right| \\
&\le \frac{|x|^{n+1}}{(n+1)!} \le |x|^{n+1}
\end{aligned} \tag{27}$$

where the first inequality holds by Lemma 1, the second inequality holds by $|\exp(\mathrm{i}s)| = 1$ as well as $x - s \le 0$ for $x \le s \le 0$ and $x - s \ge 0$ for $x \ge s \ge 0$, and the third inequality holds by direct integral calculation. Note that $\frac{x^{n+1}}{n+1} = \int_0^x (x-s)^n\, ds$, replace it into (24) and rearrange the term to get

$$\begin{aligned}
&\int_0^x (x-s)^n (\exp(\mathrm{i}s) - 1)\, ds \\
&= \frac{i}{n+1} \int_0^x (x-s)^{n+1} \exp(\mathrm{i}s)\, ds,
\end{aligned} \tag{28}$$

therefore,

$$\begin{aligned}
&\left| \frac{\mathrm{i}^{n+1}}{n!} \int_0^x (x-s)^n \exp(\mathrm{i}s)\, ds \right| \\
&= \left| \frac{\mathrm{i}^{n}}{(n-1)!} \right| \cdot \left| \frac{i}{n} \int_0^x (x-s)^n \exp(\mathrm{i}s)\, ds \right| \\
&= \frac{1}{(n-1)!} \left| \int_0^x (x-s)^{n-1} (\exp(\mathrm{i}s) - 1)\, ds \right| \\
&\le \frac{1}{(n-1)!} \int_0^x |(x-s)^{n-1} (\exp(\mathrm{i}s) - 1)|\, ds \\
&\le \frac{2}{(n-1)!} \left| \int_0^x (x-s)^{n-1}\, ds \right| \\
&\le \frac{2|x|^n}{n!} \le 2|x|^n
\end{aligned} \tag{29}$$

where the first inequality holds by Lemma 1, the second inequality holds by $|\exp(\mathrm{i}s) - 1| \le 2$ as well as $x - s \le 0$ for $x \le s \le 0$ and $x - s \ge 0$ for $x \ge s \ge 0$, and the third inequality holds by direct integral calculation. Put (26), (27), and (29) together, we have

$$\begin{aligned}
\left| \exp(\mathrm{i}x) - \sum_{m=0}^{n} \frac{(\mathrm{i}x)^m}{m!} \right| &= \left| \frac{\mathrm{i}^{n+1}}{n!} \int_0^x (x-s)^n \exp(\mathrm{i}s)\, ds \right| \\
&\le \min\{|x|^{n+1}, 2|x|^n\}.
\end{aligned} \tag{30}$$

Replacing x with $\langle \mathbf{t}, X \rangle$ in (30), and together with Lemma 1, we finally have

$$\begin{aligned}
&\left| \mathbb{E}[\exp(\mathrm{i}\langle \mathbf{t}, X \rangle)] - \mathbb{E}\left[\sum_{m=0}^{n} \frac{(\mathrm{i}\langle \mathbf{t}, X \rangle)^m}{m!} \right] \right| \\
&= \left| \mathbb{E}\left[\exp(\mathrm{i}\langle \mathbf{t}, X \rangle) - \sum_{m=0}^{n} \frac{(\mathrm{i}\langle \mathbf{t}, X \rangle)^m}{m!} \right] \right| \\
&\le \mathbb{E}\left[\left| \exp(\mathrm{i}\langle \mathbf{t}, X \rangle) - \sum_{m=0}^{n} \frac{(\mathrm{i}\langle \mathbf{t}, X \rangle)^m}{m!} \right| \right] \\
&\le \mathbb{E}[\min(|\langle \mathbf{t}, X \rangle|^{n+1}, 2|\langle \mathbf{t}, X \rangle|^n)].
\end{aligned} \tag{31}$$

And this is exactly what we claim. ■

Lemma 3 *Suppose* $r_{j,n} \in \mathbb{R}$, $\forall 1 \le j \le n, n \in \mathbb{N}_+$, $\lim_{n\to\infty} \max_{1\le j\le n} |r_{j,n}| = 0$, $\lim_{n\to\infty} \sum_{1\le j\le n} r_{j,n} = \lambda \in \mathbb{R}$, *and* $\sup_{n\in\mathbb{N}_+} \sum_{j=1}^{n} |r_{j,n}| < +\infty$, *then* $\lim_{n\to\infty} \prod_{1\le j\le n} (1 + r_{j,n}) = \exp(\lambda)$.

Proof of Lemma 3 : The hard part of the proof is the case that there may be infinite $c_{j,n} = 0$, so we must be delicate. We first do the following division:

$$\prod_{1\le j\le n} (1 + r_{j,n}) = \exp\Big(\sum_{1\le j\le n} r_{j,n} \Big) \prod_{1\le j\le n} \left(\frac{1 + r_{j,n}}{e^{r_{j,n}}} \right). \tag{32}$$

By the condition $\lim_{n\to\infty} \sum_{1\le j\le n} r_{j,n} = \lambda$ and the continuity of exponential function, we know

$$\lim_{n\to\infty} \exp\Big(\sum_{1\le j\le n} r_{j,n} \Big) = \exp(\lambda). \tag{33}$$

On the other hand, since $e = \lim_{x\to 0} (1+x)^{\frac{1}{x}}$, we have $e(1-\epsilon) \le (1+x)^{\frac{1}{x}} \le e(1+\epsilon)$, $\forall\, 1 > \epsilon > 0 \wedge |x| \le \delta_\epsilon \wedge |x| \ne 0$. By the condition $\lim_{n\to\infty} \max_{1\le j\le n} |r_{j,n}| = 0$, we can select N_ϵ s.t. $|r_{j,n}| \le \delta_\epsilon$, $\forall\, n \ge N_\epsilon \wedge 1 \le j \le n$, and if $r_{j,n} \ne 0$, then

$$\frac{1 + r_{j,n}}{e^{r_{j,n}}} = \left(\frac{(1 + r_{j,n})^{\frac{1}{r_{j,n}}}}{e} \right)^{sign(r_{j,n})|r_{j,n}|} \tag{34}$$

$$1 - \epsilon \le \left(\frac{(1 + r_{j,n})^{\frac{1}{r_{j,n}}}}{e} \right) \le 1 + \epsilon \tag{35}$$

$$\frac{1}{1+\epsilon} \le \left(\frac{(1 + r_{j,n})^{\frac{1}{r_{j,n}}}}{e} \right)^{-1} \le \frac{1}{1-\epsilon} \tag{36}$$

$$\left(\min\{1 - \epsilon, \frac{1}{1+\epsilon}\} \right)^{|r_{j,n}|} \le \frac{1 + r_{j,n}}{e^{r_{j,n}}}$$

$$\leq \left(\max\{1+\epsilon, \frac{1}{1-\epsilon}\}\right)^{|r_{j,n}|}, \tag{37}$$

and if $r_{j,n}=0$, then $\frac{1+r_{j,n}}{e^{r_{j,n}}}=1$, thus

$$1=\left(\min\{1-\epsilon, \frac{1}{1+\epsilon}\}\right)^{|r_{j,n}|} \leq \frac{1+r_{j,n}}{e^{r_{j,n}}} \leq \left(\max\{1+\epsilon, \frac{1}{1-\epsilon}\}\right)^{|r_{j,n}|}=1. \tag{38}$$

Combine the case $r_{j,n}=0$ and $r_{j,n}\neq 0$, we get

$$\left(\min\{1-\epsilon, \frac{1}{1+\epsilon}\}\right)^{\sum_{1\leq j\leq n}|r_{j,n}|} \leq \prod_{1\leq j\leq n}\left(\frac{1+r_{j,n}}{e^{r_{j,n}}}\right) \leq \left(\max\{1+\epsilon, \frac{1}{1-\epsilon}\}\right)^{\sum_{1\leq j\leq n}|r_{j,n}|}, \forall\, n\geq N_\epsilon\,. \tag{39}$$

Since we have the condition $\sup_{n\in\mathbb{N}_+}\sum_{j=1}^n |r_{j,n}| = S < +\infty$, taking the limit in (39), we derive

$$\left(\min\{1-\epsilon, \frac{1}{1+\epsilon}\}\right)^{S} \leq \liminf_{n\to\infty}\prod_{1\leq j\leq n}\left(\frac{1+r_{j,n}}{e^{r_{j,n}}}\right) \leq \limsup_{n\to\infty}\prod_{1\leq j\leq n}\left(\frac{1+r_{j,n}}{e^{r_{j,n}}}\right) \leq \left(\max\{1+\epsilon, \frac{1}{1-\epsilon}\}\right)^{S} \tag{40}$$

where $\forall\, 0<\epsilon<1$, which means

$$\lim_{n\to\infty}\prod_{1\leq j\leq n}\left(\frac{1+r_{j,n}}{e^{r_{j,n}}}\right)=1\,. \tag{41}$$

Replacing (33) and (41) in (32), we can get

$$\lim_{n\to\infty}\prod_{1\leq j\leq n}(1+r_{j,n})=\exp(\lambda)\cdot 1=\exp(\lambda), \tag{42}$$

which is the desired result. ■

Lemma 4 *Suppose* $a_i, b_i \in \mathbb{C} \wedge |a_i|\leq r \wedge |b_i|\leq r$, $\forall i\in\mathbb{N}_+$ *where* $r\in\mathbb{R}_+$ *is some positive real number, then* $|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i| \leq r^{n-1}\sum_{i=1}^n |a_i-b_i|$ *for any positive integer* n.

Proof of Lemma 4 : We prove it by induction. The case when $n=1$ is obviously true, the induction is as follows:

$$\begin{aligned}
&\left|\prod_{i=1}^{n+1}a_i - \prod_{i=1}^{n+1}b_i\right| \\
&= \left|\prod_{i=1}^{n+1}a_i - a_{n+1}\prod_{i=1}^{n}b_i + a_{n+1}\prod_{i=1}^{n}b_i - \prod_{i=1}^{n+1}b_i\right| \\
&\leq |a_{n+1}|\left|\prod_{i=1}^{n}a_i - \prod_{i=1}^{n}b_i\right| + \left|\prod_{i=1}^{n}b_i\right||a_{n+1}-b_{n+1}| \\
&\leq r^n\sum_{i=1}^n|a_i-b_i| + r^n|a_{n+1}-b_{n+1}| \leq r^n\sum_{i=1}^{n+1}|a_i-b_i|\,.
\end{aligned} \tag{43}$$

Thus the result is proved. ■

Proof of Theorem 3 (Continued) : According to the concept of the asymptotic efficiency in [4], the asymptotic distribution of people motion in a video is defined as the weak convergence limit of $M_n = \sqrt{n}(\frac{\sum_{j=1}^n \mathbf{q}_j}{n}) = \frac{\sum_{j=1}^n \mathbf{q}_j}{\sqrt{n}}$, i.e., $M_n \xrightarrow{n\to\infty} \mathcal{L}_\infty$ where $\mathbf{q}_j$ is the random movement of one person between two frames and '$\Rightarrow$' means the convergence in distribution. Since we only consider the variance of $\exp(\mathrm{i}\mathbf{m}^T\mathbf{w})$ where $\mathbf{w}$ is a fixed 2D vector and $\mathbf{m}\sim\mathcal{L}_\infty$, the mean of $\mathcal{L}_\infty$ does not matter. Thus, without loss of generality, we can assume $\mathbb{E}[\mathbf{q}_j]=0, \forall\, j\in\mathbb{N}_+$.

According to the definition of weak convergence as well as the truth that $\exp(\mathrm{i}\langle\mathbf{x},\mathbf{w}\rangle)$ is a bounded continuous function of $\mathbf{x}$ for a fixed $\mathbf{w}$, $M_n \Rightarrow \mathcal{L}_\infty$ indicates that

$$\lim_{n\to\infty}\mathbb{E}[\exp(\mathrm{i}\langle M_n,\mathbf{w}\rangle)] = \mathbb{E}_{\mathbf{m}\sim\mathcal{L}_\infty}[\exp(\mathrm{i}\langle\mathbf{m},\mathbf{w}\rangle)] \tag{44}$$

for any $\mathbf{w}\in\mathbb{R}^2$. By independence, we can write

$$\mathbb{E}[\exp(\mathrm{i}\langle M_n,\mathbf{w}\rangle)] = \prod_{j=1}^n \mathbb{E}\left[\exp(\mathrm{i}\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle)\right]\,. \tag{45}$$

Now we look at the term $\mathbb{E}\left[\exp(\mathrm{i}\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle)\right]$ where $\mathbf{w}$ is fixed and $\mathbf{q}_j$ is some random movement vector with mean $\mathbf{0}$. Using Lemma 2, we know

$$\begin{aligned}
&\left|\mathbb{E}\left[\exp(\mathrm{i}\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle)\right] - \mathbb{E}\left[\sum_{k=0}^{2}\frac{(\mathrm{i}\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle)^k}{k!}\right]\right| \\
&\leq \mathbb{E}\left[\min\left(\left|\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle\right|^3, 2\left|\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle\right|^2\right)\right] \qquad (46) \\
&\leq \mathbb{E}\left[\min\left(\left|\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle\right|^3, 2\left|\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle\right|^2\right)\cdot\mathbb{1}\{\|\mathbf{q}_j\|_2\leq\sqrt{n}\epsilon\}\right] \\
&+ \mathbb{E}\left[\min\left(\left|\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle\right|^3, 2\left|\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle\right|^2\right)\cdot\mathbb{1}\{\|\mathbf{q}_j\|_2>\sqrt{n}\epsilon\}\right] \qquad (47) \\
&\leq \mathbb{E}\left[\left|\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle\right|^3\cdot\mathbb{1}\{\|\mathbf{q}_j\|_2\leq\sqrt{n}\epsilon\}\right] \\
&\quad + \mathbb{E}\left[2\left|\langle\frac{\mathbf{q}_j}{\sqrt{n}},\mathbf{w}\rangle\right|^2\cdot\mathbb{1}\{\|\mathbf{q}_j\|_2>\sqrt{n}\epsilon\}\right] \qquad (48)
\end{aligned}$$

$$\leq \frac{1}{\sqrt{n}^3}\mathbb{E}\left[\left|\langle \mathbf{q}_j, \mathbf{w}\rangle\right|^2 \|\mathbf{q}_j\|_2\|\mathbf{w}\|_2 \cdot \mathbb{1}\{\|\mathbf{q}_j\|_2 \leq \sqrt{n}\epsilon\}\right] + \frac{2}{n}\mathbb{E}\left[\left|\langle \mathbf{q}_j, \mathbf{w}\rangle\right|^2 \cdot \mathbb{1}\{\|\mathbf{q}_j\|_2 > \sqrt{n}\epsilon\}\right] \tag{49}$$

$$\leq \frac{\epsilon\|\mathbf{w}\|_2}{n}\mathbb{E}\left[\langle \mathbf{q}_j, \mathbf{w}\rangle^2\right] + \frac{2}{n}\mathbb{E}\left[\langle \mathbf{q}_j, \mathbf{w}\rangle^2 \cdot \mathbb{1}\{\|\mathbf{q}_j\|_2 > \sqrt{n}\epsilon\}\right] \tag{50}$$

where (46) holds by Lemma 2, (47) holds by dividing the integral region, and (49) holds by Cauchy-Schwartz inequality. Fixing $\mathbf{w}$, $\tilde{q}_j^{\mathbf{w}} = \langle \mathbf{q}_j, \mathbf{w}\rangle$ is a random variable, and

$$\mathbb{E}[\tilde{q}_j^{\mathbf{w}}] = \langle \mathbb{E}[\mathbf{q}_j], \mathbf{w}\rangle = 0, \tag{51}$$

$$\mathbb{E}[(\tilde{q}_j^{\mathbf{w}})^2] = \mathrm{var}(\tilde{q}_j^{\mathbf{w}}) + \mathbb{E}[\tilde{q}_j^{\mathbf{w}}]^2 = \mathbf{w}^T\mathrm{cov}(\mathbf{q}_j)\mathbf{w} + 0 = \mathbf{w}^T\mathrm{cov}(\mathbf{q}_j)\mathbf{w}\,. \tag{52}$$

Note that the image size of each frame is bounded, thus each random movement $\mathbf{q}_j$ is also bounded, therefore $(\tilde{q}_j^{\mathbf{w}})^2$ is bounded by some real number $B_{\mathbf{w}}$ for each fixed $\mathbf{w}$ and all $j \in \mathbb{N}_+$. Hence, together with (50), we have

$$\sum_{j=1}^n \left|\mathbb{E}\left[\exp(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)\right] - \mathbb{E}\left[\sum_{k=0}^2 \frac{(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)^k}{k!}\right]\right| \leq \frac{\epsilon\|\mathbf{w}\|_2}{n}\mathbf{w}^T\sum_{j=1}^n \mathrm{cov}(\mathbf{q}_j)\mathbf{w} + 2B_{\mathbf{w}}\mathbb{E}[\mathbb{1}\{\|\mathbf{q}_j\|_2 > \sqrt{n}\epsilon\}]$$

$$\leq \epsilon\|\mathbf{w}\|_2\mathbf{w}^T\mathrm{cov}(M_n)\mathbf{w} + 2B_{\mathbf{w}}\mathbb{E}[\mathbb{1}\{\|\mathbf{q}_j\|_2 > \sqrt{n}\epsilon\}]\,. \tag{53}$$

According to the condition in the theorem, $\mathbf{\Lambda}$ is the covariance matrix of $\mathcal{L}_\infty$. Because $M_n \Rightarrow \mathcal{L}_\infty$ as $n \to \infty$, and $\|\mathrm{cov}(M_n)\|_F$ (the Frobenius norm of matrix) is bounded since each movement is uniformly bounded by the image size, we have $\mathrm{cov}(M_n) \to \mathbf{\Lambda}$ as $n \to \infty$. Also, by dominated convergence theorem, $\mathbb{E}[\mathbb{1}\{\|\mathbf{q}_j\|_2 > \sqrt{n}\epsilon\}] \to 0$ as $n \to \infty$. Therefore, together with (53), we have

$$\limsup_{n\to\infty} \sum_{j=1}^n \left|\mathbb{E}\left[\exp(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)\right] - \mathbb{E}\left[\sum_{k=0}^2 \frac{(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)^k}{k!}\right]\right| \leq \epsilon\|\mathbf{w}\|_2\mathbf{w}^T\mathbf{\Lambda}\mathbf{w},\ \forall\,\epsilon > 0\,. \tag{54}$$

Since ϵ can be arbitrarily small, and $\|\mathbf{w}\|_2\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}$ is constant for fixed $\mathbf{w}$, we in fact have

$$\lim_{n\to\infty} \sum_{j=1}^n \left|\mathbb{E}\left[\exp(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)\right] - \mathbb{E}\left[\sum_{k=0}^2 \frac{(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)^k}{k!}\right]\right| = 0\,. \tag{55}$$

On the other hand, by (51) and (52),

$$\mathbb{E}\left[\sum_{k=0}^2 \frac{(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)^k}{k!}\right] = 1 + \frac{\mathrm{i}}{\sqrt{n}}\mathbb{E}[\tilde{q}_j^{\mathbf{w}}] - \frac{1}{2n}\mathbb{E}[(\tilde{q}_j^{\mathbf{w}})^2] = 1 - \frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2n}\,. \tag{56}$$

When n is large enough, $|1 - \frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2n}| \leq 1$, and by Lemma 1, $\left|\mathbb{E}\left[\exp(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)\right]\right| \leq \mathbb{E}\left[\left|\exp(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)\right|\right] = \mathbb{E}[1] = 1$. Hence taking $r = 1$ in Lemma 4, we can apply it with (45), (55), and (56) to get

$$\lim_{n\to\infty}\left|\mathbb{E}[\exp(\mathrm{i}\langle M_n, \mathbf{w}\rangle)] - \prod_{j=1}^n\left(1 - \frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2n}\right)\right| \leq \lim_{n\to\infty}\sum_{j=1}^n\left|\mathbb{E}\left[\exp(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)\right] - \mathbb{E}\left[\sum_{k=0}^2 \frac{(\mathrm{i}\langle \frac{\mathbf{q}_j}{\sqrt{n}}, \mathbf{w}\rangle)^k}{k!}\right]\right| \leq 0\,. \tag{57}$$

Note that $\sum_{j=1}^n -\frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2n} = -\frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2}$ and $-\frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2n} \to 0$ as $n \to \infty$, therefore by Lemma 3 we have

$$\lim_{n\to\infty}\prod_{j=1}^n\left(1 - \frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2n}\right) = \exp\left(-\frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2}\right)\,. \tag{58}$$

(44), (57), and (58) together show that

$$\mathbb{E}_{\mathbf{m}\sim\mathcal{L}_\infty}[\exp(\mathrm{i}\mathbf{m}^T\mathbf{w})] = \exp\left(-\frac{\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}}{2}\right)\,. \tag{59}$$

Therefore, we have

$$\mathrm{var}_{\mathbf{m}_j\sim\mathcal{L}_\infty}(\exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w})) = \mathbb{E}_{\mathbf{m}_j\sim\mathcal{L}_\infty}[|\exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w})|^2] - |\mathbb{E}_{\mathbf{m}_j\sim\mathcal{L}_\infty}[\exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w})]|^2 \tag{60}$$

$$= 1 - |\mathbb{E}_{\mathbf{m}_j\sim\mathcal{L}_\infty}[\exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w})]|^2 \tag{61}$$

$$= 1 - \exp\left(-\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}\right) \tag{62}$$

where (61) holds because that $|\exp(\mathrm{i}\mathbf{m}_j^T\mathbf{w})|^2 = 1$, and (62) holds by (59). Substituting it into (20), we have

$$\mathrm{std}_{\mathbf{m}_j\sim\mathcal{N}(0,\mathbf{\Lambda})}[\Delta(\mathbf{w})] = \sqrt{Q(1 - \exp(-\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}))\exp(-\mathbf{w}^T\mathbf{\Sigma}\mathbf{w})}\,. \tag{63}$$

■

C.4. Proof of Theorem 4

We prove a more sophisticated case, in which the theorem is just one special setting in the more general case.

We start by clarifying a concept that has been repeatedly used and some symbols.

Definition 1 *Suppose we have an entire set Ω, $\mathcal{A}$ is a family of subsets of Ω. We call $\mathcal{A}$ a* ***σ-algebra*** *if*

- $X \in \mathcal{A}$;
- $A_i \in \mathcal{A}, i \in \mathbb{N}_+ \Rightarrow \cup_{i=1}^{+\infty} A_i \in \mathcal{A}$;
- $A \in \mathcal{A} \Rightarrow X - A \in \mathcal{A}$.

The second and third conditions easily infer the closeness under countable intersection, thus σ-algebra is a family of set closed under countable union, countable intersection, and complement, as well as including the whole set.

*Suppose $\mathcal{A}$ is a family of subsets of Ω, then $\sigma(\mathcal{A})$ is the smallest σ-algebra containing $\mathcal{A}$, i.e. $\sigma(\mathcal{A})$ is a σ-algebra, and for any other σ-algebra $\mathcal{H}$ where $\mathcal{H} \supset \mathcal{A}$, we have $\sigma(\mathcal{A}) \subset \mathcal{H}$. $\sigma(\mathcal{A})$ is called the **generated σ-algebra** by $\mathcal{A}$.*

Suppose X is a function from a space Ω to another measurable space $(S, \mathcal{S})$, then $\sigma(X)$ is the smallest σ-algebra making X measurable, i.e.,

$$\sigma(X) = \{X^{-1}(A) \mid \forall\, A \in \mathcal{S}\},$$

*and it's called the **generated σ-algebra** by X.*

Suppose there is an index set J, for each $j \in J$, X_j is a function from a space Ω to another measurable space $(S_j, \mathcal{S}_j)$, then $\sigma(\cup_{j\in J} X_j)$ is the smallest σ-algebra making each X_j measurable, i.e.,

$$\sigma(\cup_{j\in J} X_j) = \sigma(\cup_{j\in J} \sigma(X_j)),$$

*and it's called the **generated σ-algebra** by $\{X_j, j \in J\}$. If J is a finite set with its elements $j_1, j_2, \ldots, j_n$, we also write $\sigma(\cup_{j\in J} X_j)$ as $\sigma(X_{j_1}, \ldots, X_{j_n})$.*

Now we build up the basic setting based on the above definition. Suppose $\mathcal{F}$ is the σ-algebra of all the information of the video $\mathcal{V} = (\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n)$, i.e., $\mathcal{F} = \sigma(\mathcal{I}_1, \cdots, \mathcal{I}_n)$. The current time point is t, we may want to predict (infer) some square-integrable information at time t, e.g. the total number of people, the position of each person, etc. All of that information composes a random vector $W^{(t)}$ that $\mathbb{E}[\|W^{(t)}\|_2^2] < +\infty$, where each dimension of $W^{(t)}$ corresponds to a specific square-integrable value such as the total number of people, the x-coordinate of person one, the y-coordinate of person one, the density map value at some coordinate (x', y'), etc. The range space of each frame $\mathcal{I}_t$, $1 \le t \le n$ is the Borel tensor space $R_{\mathcal{I}} = (\mathbb{R}^{c\times h\times w}, \mathcal{B}_{\mathbb{R}^{c\times h\times w}})$. Now, we need to extend the definition of conditional expectation to multi-dimension.

Definition 2 *For a k-dimensional ($k \in \mathbb{N}_+$) random vector W and a σ-algebra $\mathcal{H}$, we define the **conditional expectation** of W w.r.t. $\mathcal{H}$ as the random vector $\mathbb{E}[W|\mathcal{H}] = [\mathbb{E}[W_1|\mathcal{H}], \mathbb{E}[W_2|\mathcal{H}], \cdots, \mathbb{E}[W_k|\mathcal{H}]]$*

The next definition is the clarification of some symbols.

Definition 3 *For a k-dimensional ($k \in \mathbb{N}_+$) random vector W, $\mathbb{E}[W|\mathcal{V}] = \mathbb{E}[W|\sigma(\mathcal{V})] = \mathbb{E}[W|\mathcal{F}]$, and $\mathbb{E}[W|\mathcal{I}_{j_1}, \mathcal{I}_{j_2}, \cdots, \mathcal{I}_{j_m}] = \mathbb{E}[W|\sigma(\mathcal{I}_{j_1}, \mathcal{I}_{j_2}, \cdots, \mathcal{I}_{j_m})]$, i.e., the conditional expectation of W w.r.t. the smallest σ-algebra making $\mathcal{I}_{j_1}, \mathcal{I}_{j_2}, \cdots, \mathcal{I}_{j_m}$ measurable.*

In preparation for proving the final theorem, we begin by deriving a set of auxiliary lemmas that progressively construct the theoretical foundation needed for the complete proof.

Lemma 5 *Let $\mathcal{A}$ be a family of subsets of S, suppose X is a function from Ω to $(S, \sigma(\mathcal{A}))$ where $\sigma(\mathcal{A})$ means the generated σ-algebra of $\mathcal{A}$, then $\sigma(X) = \sigma(X^{-1}(\mathcal{A}))$ where $X^{-1}(\mathcal{A}) = \{X^{-1}(A) \mid \forall\, A \in \mathcal{A}\}$.*

Proof of Lemma 5 : By Definition 1, $\sigma(X)$ is the smallest σ-algebra s.t. X measurable, i.e.,

$$\sigma(X) = \{X^{-1}(A) \mid \forall\, A \in \sigma(\mathcal{A})\}, \tag{64}$$

then obviously $X^{-1}(\mathcal{A}) \subset \sigma(X)$ since $\mathcal{A} \subset \sigma(\mathcal{A})$, then

$$\sigma(X^{-1}(\mathcal{A})) \subset \sigma(X) \tag{65}$$

because $\sigma(X^{-1}(\mathcal{A}))$ is the smallest σ-algebra contains $X^{-1}(\mathcal{A})$.

For the other side, we define $\mathcal{D} = \{A \in \sigma(\mathcal{A}) : X^{-1}(A) \in \sigma(X^{-1}(\mathcal{A}))\}$, then

$$\mathcal{A} \subset \mathcal{D} \subset \sigma(\mathcal{A})\,. \tag{66}$$

Now we claim $\mathcal{D}$ is an algebra (see the definition in Item 3 of Section 1.5 in [7]) because

- $X^{-1}(S) = \Omega \in \sigma(X^{-1}(\mathcal{A})) \Rightarrow S \in \mathcal{D}$;
- $$A_1, A_2 \in \mathcal{D} \Rightarrow$$ $$A_1 \cap A_2 \in \sigma(\mathcal{A}) \wedge$$ $$X^{-1}(A_1 \cap A_2) = X^{-1}(A_1) \cap X^{-1}(A_2) \in \sigma(X^{-1}(\mathcal{A}))$$ $$\Rightarrow A_1 \cap A_2 \in \mathcal{D};$$
- $$A_1, A_2 \in \mathcal{D} \Rightarrow$$ $$A_1 \cup A_2 \in \sigma(\mathcal{A}) \wedge$$ $$X^{-1}(A_1 \cup A_2) = X^{-1}(A_1) \cup X^{-1}(A_2) \in \sigma(X^{-1}(\mathcal{A}))$$ $$\Rightarrow A_1 \cup A_2 \in \mathcal{D};$$
- $$A \in \mathcal{D} \Rightarrow$$ $$S - A \in \sigma(\mathcal{A}) \wedge$$ $$X^{-1}(S - A) = \Omega - X^{-1}(A) \in \sigma(X^{-1}(\mathcal{A}))$$ $$\Rightarrow S - A \in \mathcal{D}.$$

We also claim that $\mathcal{D}$ is a monotone class (see the definition in Item 6 of Section 1.5 in [7]) because

- $$A_i \subset A_{i+1}, A_i \in \mathcal{D}, \forall i \in \mathbb{N}_+ \Rightarrow$$
$$\bigcup_{i=1}^{+\infty} A_i \in \sigma(\mathcal{A}) \wedge$$
$$X^{-1}\left(\bigcup_{i=1}^{+\infty} A_i\right) = \bigcup_{i=1}^{+\infty} X^{-1}(A_i) \in \sigma(X^{-1}(\mathcal{A}))$$
$$\Rightarrow \bigcup_{i=1}^{+\infty} A_i \in \mathcal{D};$$
- $$A_i \supset A_{i+1}, A_i \in \mathcal{D}, \forall i \in \mathbb{N}_+ \Rightarrow$$
$$\bigcap_{i=1}^{+\infty} A_i \in \sigma(\mathcal{A}) \wedge$$
$$X^{-1}\left(\bigcap_{i=1}^{+\infty} A_i\right) = \bigcap_{i=1}^{+\infty} X^{-1}(A_i) \in \sigma(X^{-1}(\mathcal{A}))$$
$$\Rightarrow \bigcap_{i=1}^{+\infty} A_i \in \mathcal{D}.$$

Thus by the monotone class theorem (Theorem 1.5 of Section 1.5 in [7]), $\mathcal{D}$ is a σ-algebra. Since $\mathcal{D}$ also contains $\mathcal{A}$ by its definition, we have

$$\mathcal{D} \supset \sigma(\mathcal{A}) \tag{67}$$

because $\sigma(\mathcal{A})$ is the smallest σ-algebra contains $\mathcal{A}$ by Definition 1. (67) and (66) show $\mathcal{D} = \sigma(\mathcal{A})$, which means

$$X^{-1}(A) \in \sigma(X^{-1}(\mathcal{A})),\ \forall A \in \sigma(\mathcal{A}), \tag{68}$$

therefore by (64),

$$\sigma(X) \subset \sigma(X^{-1}(\mathcal{A}))\ . \tag{69}$$

(65) and (69) show that $\sigma(X) = \sigma(X^{-1}(\mathcal{A}))$. ■

Lemma 6 *Suppose J is an index set, $(\Omega, \mathcal{H})$, $(S_j, \mathcal{S}_j), \forall j \in J$ are all measurable spaces, and their product space is denoted by $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$. And for any $j \in J$, X_j is a function from Ω to $(S_j, \mathcal{S}_j)$, then $\prod_{j\in J} X_j$ is a function from Ω to $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$. Then we have $\sigma(\cup_{j\in J} X_j) = \sigma(\prod_{j\in J} X_j)$.*

Proof of Lemma 6 : We first define a family of set

$$\mathcal{A} = \{A_i \times \prod_{j\neq i, j\in J} S_j \mid A_i \in \mathcal{S}_i\ ,\ i \in J\}\ . \tag{70}$$

According to the definition of product σ-algebra (see Definition 6.18 in [2]),

$$\prod_{j\in J} \mathcal{S}_j = \sigma(\mathcal{A})\ . \tag{71}$$

Thus by Lemma 5,

$$\sigma(\prod_{j\in J} X_j) = \sigma((\prod_{j\in J} X_j)^{-1}(\mathcal{A}))\ . \tag{72}$$

Since $(\prod_{j\in J} X_j)^{-1}(A_i \times \prod_{j\neq i, j\in J} S_j) = X_i^{-1}(A_i)$, we have

$$(\prod_{j\in J} X_j)^{-1}(\mathcal{A}) = \bigcup_{j\in J} \sigma(X_j)\ . \tag{73}$$

Now, (72), (73), and Definition 1 show that

$$\sigma(\prod_{j\in J} X_j) = \sigma(\bigcup_{j\in J} \sigma(X_j)) = \sigma(\cup_{j\in J} X_j)\ . \tag{74}$$

■

Lemma 7 *Suppose J is an index set, $(\Omega, \mathcal{H})$, $(S_j, \mathcal{S}_j), \forall j \in J$ are all measurable spaces, and their product space is denoted by $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$. And for any $j \in J$, X_j is a function from Ω to $(S_j, \mathcal{S}_j)$, then $\prod_{j\in J} X_j$ is a function from Ω to $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$. We have $\prod_{j\in J} X_j$ is measurable from $(\Omega, \mathcal{H})$ to $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$ iff each X_j is measurable from $(\Omega, \mathcal{H})$ to $(S_j, \mathcal{S}_j)$.*

Proof of Lemma 7 : Suppose each X_j is measurable from $(\Omega, \mathcal{H})$ to $(S_j, \mathcal{S}_j)$, which means $\mathcal{H} \supset \sigma(X_j)$ for any $j \in J$. Therefore $\mathcal{H} \supset \sigma(\cup_{j\in J} X_j)$, and by Lemma 6 we have

$$\mathcal{H} \supset \sigma(\prod_{j\in J} X_j), \tag{75}$$

which means $\prod_{j\in J} X_j$ is measurable from $(\Omega, \mathcal{H})$ to $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$.

Now suppose $\prod_{j\in J} X_j$ is measurable from $(\Omega, \mathcal{H})$ to $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$, which means $\mathcal{H} \supset \sigma(\prod_{j\in J} X_j)$. By Lemma 6, it means $\mathcal{H} \supset \sigma(\cup_{j\in J} X_j)$, which by Definition 1 indicates that

$$\mathcal{H} \supset \sigma(X_j),\ \forall\ j \in J\ . \tag{76}$$

And this shows that each X_j is measurable on $(\Omega, \mathcal{H})$. ■

The next Lemmas need the definition of simple functions.

Definition 4 *$\mathbb{1}_A(x)$ is called an **indicator function** if A is a measurable set and $\mathbb{1}_A(x) = 1$ when $x \in A$ and otherwise 0. a **simple function** is the linear combination of finite indicator functions.*

Lemma 8 *Suppose f is a measurable function from $(\Omega, \mathcal{H})$ to $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ where $\mathcal{B}_\mathbb{R}$ means all the Borel sets of $\mathbb{R}$, then there is a series of simple functions $\{f_n\}_{n\in\mathbb{N}_+}$ from $(\Omega, \mathcal{H})$ to $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ s.t. $f(x) = \lim_{n\to\infty} f_n(x)$ for any $x \in \Omega$.*

Proof of Lemma 8 : Define

$$f_n(x) = \begin{cases} -n, & f(x) \le -n \\ -n + \frac{kn}{2^n}, & -n + \frac{kn}{2^n} \le f(x) < -n + \frac{(k+1)n}{2^n}, \\ & 0 \le k \le 2^{n+1} - 1 \\ n, & f(x) \ge n \end{cases} \tag{77}$$

Since f is measurable function from $(\Omega, \mathcal{H})$ to $(\mathbb{R}, \mathcal{B}_\mathbb{R})$, we can define the following measurable set of $(\Omega, \mathcal{H})$ for $n \in \mathbb{N}_+$:

$$A_0^n = \{x \in \Omega : f(x) < -n + \frac{n}{2^n}\}; \tag{78}$$

$$A_k^n = \{x \in \Omega : -n + \frac{kn}{2^n} \le f(x) < -n + \frac{(k+1)n}{2^n}\}, \quad \forall\, 1 \le k \le 2^{n+1} - 1; \tag{79}$$

$$A_{2^{n+1}}^n = \{x \in \Omega : f(x) \ge n\}\,. \tag{80}$$

Then we claim $f_n(x)$ is a simple function because $f_n(x) = n\mathbb{1}_{A_{2^{n+1}}^n}(x) + \sum_{k=0}^{2^{n+1}-1}(-n + \frac{kn}{2^n})\mathbb{1}_{A_k^n}(x)$ is the linear combination of finite indicator functions. Moreover, for any $x \in \Omega$, we can always find large enough N s.t. $\forall\, n \ge N$, $-n + \frac{k^{(n)}n}{2^n} \le f(x) < -n + \frac{(k^{(n)}+1)n}{2^n}$ for a unique $0 \le k^{(n)} \le 2^{n+1} - 1$. Let $B_{k^{(n)}} = [-n + \frac{k^{(n)}n}{2^n}, -n + \frac{(k^{(n)}+1)n}{2^n})$, Since $f(x) \in B_{k^{(n)}}$ for each $n \ge N$, and the inclusion interval is uniformly bisected with each unitary increment of n, we have

$$\forall\, n \ge N,\ B_{k^{(n+1)}} \subset B_{k^{(n)}}\,. \tag{81}$$

Since each such interval includes $f(x)$ and the length of such intervals tends to zero as n tends to infinity, it shows

$$-n + \frac{k^{(n)}n}{2^n} \xrightarrow{n\to\infty} f(x), \tag{82}$$

which implies $\lim_{n\to\infty} f_n(x) = f(x)$. ■

Lemma 9 *Suppose J is an index set, $(S_j, \mathcal{S}_j), \forall j \in J$ are all measurable spaces. For any $j \in J$, X_j is a function from Ω to $(S_j, \mathcal{S}_j)$, and $\prod_{j\in J} X_j$ is a function from Ω to the product space $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$. $\mathcal{B}_{\mathbb{R}^k}$ is the Borel σ-algebra of $\mathbb{R}^k$. Then, for any $k \in \mathbb{N}_+$, Y is measurable from $(\Omega, \sigma(\cup_{j\in J} X_j))$ to $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$ iff there is a function f measurable from $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$ to $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$ s.t. $Y = f(\prod_{j\in J} X_j)$.*

Proof of Lemma 9 : First suppose f is measurable from $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$ to $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$ and $Y = f(\prod_{j\in J} X_j)$, then $\forall A \in \mathcal{B}_{\mathbb{R}^k}$,

$$Y^{-1}(A) = (\prod_{j\in J} X_j)^{-1}(f^{-1}(A))\,. \tag{83}$$

Because f is measurable, $f^{-1}(A) \in \prod_{j\in J} \mathcal{S}_j$, thus $(\prod_{j\in J} X_j)^{-1}(f^{-1}(A)) \in \sigma(\prod_{j\in J} X_j)$ by Definition 1. By Lemma 6, it shows $Y^{-1}(A) \in \sigma(\cup_{j\in J} X_j)$ for any $A \in \mathcal{B}_{\mathbb{R}^k}$, which indicates that Y is measurable from $(\Omega, \sigma(\cup_{j\in J} X_j))$ to $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$.

For the other side, we first prove the case when $k = 1$. If Y is an indicator function, then Y can be written as $Y = \mathbb{1}_A(x)$ where $A \in \sigma(\cup_{j\in J} X_j)$. By Lemma 6, $A \in \sigma(\prod_{j\in J} X_j)$. By Definition 1, there is a set $B \in \mathcal{B}_\mathbb{R}$ s.t. $A = (\prod_{j\in J} X_j)^{-1}(B)$. Let $f = \mathbb{1}_B(x)$, then $Y = f(\prod_{j\in J} X_j)$.

If Y is a simple function, then by Definition 4 Y can be written as $Y = \sum_{i=1}^n c_i \mathbb{1}_{A_i}(x)$. $\mathbb{1}_{A_i}$ can be written as $f_i(\prod_{j\in J} X_j)$ by the previous analysis, where f_i is some indicator function from $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$ to $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. Then $f = \sum_{i=1}^n c_i f_i$ is a measurable function from $(\prod_{j\in J} S_j, \prod_{j\in J} \mathcal{S}_j)$ to $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ because it's the linear combination of finite measurable functions (see Theorem 5.3 in [7]), and $Y = f(\prod_{j\in J} X_j)$.

If Y is a general measurable function, then by Lemma 8, Y can be written as $Y = \lim_{i\to+\infty} Y_i$ where each Y_i is a simple function. By the previous analysis, $Y_i = f_i(\prod_{j\in J} X_j)$ for some measurable f_i. Thus $Y = \lim_{n\to+\infty} f_i(\prod_{j\in J} X_j)$. Let $f(x) = \limsup_{i\to+\infty} f_i(x)$, then $f(x)$ is measurable as each f_i is measurable (see Theorem 5.4 in [7]), and $Y = f(\prod_{j\in J} X_j)$.

Next we prove the case when $k > 1$. We can write Y as $Y = (Y_1, Y_2, \ldots, Y_k)$ and each Y_i $(1 \le i \le k)$ is measurable from $(\Omega, \sigma(\cup_{j\in J} X_j))$ to $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ by Lemma 7. Then by applying the proved case when $k = 1$ to each Y_i, we can write

$$Y = (f_1(\prod_{j\in J} X_j), \ldots, f_k(\prod_{j\in J} X_j)) \tag{84}$$

where each f_i is measurable from $(\Omega, \sigma(\cup_{j\in J} X_j))$ to $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. By Lemma 7 again, $f = (f_1, \ldots, f_k)$ is measurable from $(\Omega, \sigma(\cup_{j\in J} X_j))$ to $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$, and $Y = f(\prod_{j\in J} X_j)$. ■

Lemma 9 offers the half of the key for proving generalized Theorem 4. For the other half key, we first need the following two lemmas w.r.t. the inner product space.

Lemma 10 *Suppose V is an inner product space over the field Λ ($\Lambda = \mathbb{R}$ or $\Lambda = \mathbb{C}$) and $\|.\|$ is the norm induced by the inner product. M is a subspace of V. Then $\forall\, x \in V, y \in M$, $\|x-y\| = \inf_{z\in M}\|x-z\|$ iff $(x-y) \perp M$.*

Proof of Lemma 10 : First suppose $\|x-y\| = \inf_{z\in M}\|x-z\|$. Let $z \in M, \lambda \in \Lambda$, then $y + \lambda z \in M$, thus

$$\|x - y - \lambda z\|^2 \geq \|x-y\|^2, \tag{85}$$

which implies

$$|\lambda|^2\|z\|^2 - 2\Re\langle x-y, \lambda z\rangle \geq 0\,. \tag{86}$$

Suppose $z \neq \mathbf{0}$, and replace λ with $\frac{\langle x-y,z\rangle}{\|z\|^2}$ in (86):

$$\begin{aligned}
&\frac{|\langle x-y,z\rangle|^2}{\|z\|^4}\|z\|^2 - 2\Re\langle x-y, \frac{\langle x-y,z\rangle}{\|z\|^2}z\rangle \\
&= \frac{|\langle x-y,z\rangle|^2}{\|z\|^2} - 2\Re(\frac{\overline{\langle x-y,z\rangle}}{\|z\|^2}\langle x-y,z\rangle) \\
&= \frac{|\langle x-y,z\rangle|^2}{\|z\|^2} - 2\frac{|\langle x-y,z\rangle|^2}{\|z\|^2} \\
&= -\frac{|\langle x-y,z\rangle|^2}{\|z\|^2} \geq 0,
\end{aligned} \tag{87}$$

which implies $\langle x-y, z\rangle = 0,\ z \in M, z \neq \mathbf{0}$. If $z = \mathbf{0}$, it's trivial $\langle x-y,z\rangle = 0$, thus we proved $x - y \perp M$.

For the other side, suppose $x - y \perp M$, then $\forall z \in M$, we have

$$\begin{aligned}
\|x-z\|^2 &= \|x-y\|^2 + \|y-z\|^2 + 2\Re\langle x-y, y-z\rangle \\
&= \|x-y\|^2 + \|y-z\|^2
\end{aligned} \tag{88}$$

where the last equality holds because $y-z \in M$ and $x-y \perp M$. (88) shows that

$$\|x-y\|^2 \leq \|x-z\|^2, \tag{89}$$

which means $\|x-y\| \leq \inf_{z\in M}\|x-z\|$. Also, $\|x-y\| \geq \inf_{z\in M}\|x-z\|$ because $y \in M$. Therefore, $\|x-y\| = \inf_{z\in M}\|x-z\|$. ■

Lemma 11 *Suppose V is an inner product space over the field Λ ($\Lambda = \mathbb{R}$ or $\Lambda = \mathbb{C}$) and $\|.\|$ is the norm induced by the inner product. M is a subspace of V. $\forall\, x \in V$, $y_1, y_2 \in M$, if $x - y_1 \perp M$ and $x - y_2 \perp M$, then $y_1 = y_2$.*

Proof of Lemma 11 : Define the orthogonal space

$$M^\perp = \{x \in V : \langle x, y\rangle = 0\ \forall y \in M\}. \tag{90}$$

Then by the linearity of the inner product, $M^\perp$ is closed under linear combination, thus a subspace. Obviously, $M \cap M^\perp = \{\mathbf{0}\}$ by the positive definiteness of the inner product. Suppose $z_1 = x - y_1$, $z_2 = x - y_2$, then

$$\begin{aligned}
&x = y_1 + z_1 = y_2 + z_2 \\
&\Rightarrow M \ni y_1 - y_2 = z_2 - z_1 \in M^\perp &(91)\\
&\Rightarrow y_1 - y_2 = 0 \Rightarrow y_1 = y_2\,. &(92)
\end{aligned}$$

■

For convenience, we introduce the following notation before proving the next lemma.

Definition 5 *Suppose we have a domain space Ω and the range space $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$ where k is some positive integer number. Then for a σ-algebra $\mathcal{H}$ of Ω, we define*

$$M_k(\mathcal{H}) = \{Y : Y \text{ measurable from } (\Omega, \mathcal{H}) \text{ to } (\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})\}. \tag{93}$$

Lemma 12 *Suppose W is a k-dimensional random vector (thus measurable) from a probability space $(\Omega, \mathcal{H}, P)$ to $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$ where k is any positive integer number and $\mathcal{B}_{\mathbb{R}^k}$ is the Borel σ-algebra of $\mathbb{R}^k$, $\mathbb{E}_P[\|W\|_2^2] < +\infty$, and $\mathcal{S} \subset \mathcal{H}$ is a sub-σ-algebra, then we have*

$$\underset{Y\in M_k(\mathcal{S})}{\arg\min}\ \mathbb{E}_P[\|W-Y\|_2^2] = \mathbb{E}[W|\mathcal{S}]\ a.s.$$

Proof of Lemma 12 : Define

$$L_k^2(\mathcal{H}) = \{Y \in M_k(\mathcal{H}) : \mathbb{E}_P[\|Y\|_2^2] < +\infty\}; \tag{94}$$
$$L_k^2(\mathcal{S}) = \{Y \in M_k(\mathcal{S}) : \mathbb{E}_P[\|Y\|_2^2] < +\infty\}. \tag{95}$$

For $X, Z \in L_k^2(\mathcal{H})$, also define

$$[X, Z] = \mathbb{E}_P[\langle X, Z\rangle]\,, \tag{96}$$

then $[.,.]$ is an inner product defined in $L_k^2(\mathcal{H})$ over the field $\mathbb{R}$ because

- for $X \in L_k^2(\mathcal{H})$:

$$\begin{aligned}
&[X,X] = 0 \iff \mathbb{E}_P[\langle X, X\rangle] = 0 \\
&\iff \mathbb{E}_P[\sum_{i=1}^k X_i^2] = \sum_{i=1}^k \mathbb{E}_P[X_i^2] = 0 \\
&\iff \mathbb{E}_P[X_i^2] = 0, \forall\, 1 \leq i \leq k \\
&\iff X_i = 0\ a.s.\ \forall\, 1 \leq i \leq k \iff X = \mathbf{0}\ a.s.\,;
\end{aligned}$$

- for $a, b \in \mathbb{R}$ and $X, Y, Z \in L_k^2(\mathcal{H})$:

$$\begin{aligned}
&[aX + bY, Z] = \mathbb{E}_P[\langle aX + bY, Z\rangle] \\
&= \mathbb{E}_P[a\langle X, Z\rangle + b\langle Y, Z\rangle] \\
&= a\mathbb{E}_P[\langle X, Z\rangle] + b\mathbb{E}_P[\langle Y, Z\rangle] \\
&= a[X,Z] + b[Y,Z];
\end{aligned}$$

- for $X, Z \in L_k^2(\mathcal{H})$:

$$[X, Z] = \mathbb{E}_P[\langle X, Z\rangle] = \mathbb{E}_P[\langle Z, X\rangle] = [Z, X].$$

Thus $L_k^2(\mathcal{H})$ equipped with the binary operation $[.,.]$ (defined in (96)) is an inner product space, and the induced norm is

$$\|Z\| = \sqrt{\mathbb{E}_P[\|Z\|_2^2]}. \tag{97}$$

Under the setting, $L_k^2(\mathcal{S})$ is a subspace because

- $L_k^2(\mathcal{S}) \subset L_k^2(\mathcal{H})$ as $M_k(\mathcal{S}) \subset M_k(\mathcal{H})$.
- By Theorem 5.3 in [7], $X, Z \in M_k(\mathcal{S}) \Rightarrow X + Z \in M_k(\mathcal{S})$. And $\mathbb{E}_P[\|X\|_2^2] < +\infty$, $\mathbb{E}_P[\|Z\|_2^2] < +\infty$ can infer

$$\begin{aligned}
&\mathbb{E}_P[\|X + Z\|_2^2] \\
&= \mathbb{E}_P[\|X\|_2^2] + \mathbb{E}_P[\|Z\|_2^2] + 2[X, Z] && (98)\\
&\le \mathbb{E}_P[\|X\|_2^2] + \mathbb{E}_P[\|Z\|_2^2] + 2\|X\|\|Z\| && (99)\\
&\le \mathbb{E}_P[\|X\|_2^2] + \mathbb{E}_P[\|Z\|_2^2] + 2\sqrt{\mathbb{E}_P[\|X\|_2^2]\,\mathbb{E}_P[\|Z\|_2^2]} && (100)\\
&< +\infty && (101)
\end{aligned}$$

  where (99) holds by Cauchy-Schwartz inequality and (100) holds by (97). Thus $X, Z \in L_k^2(\mathcal{S}) \Rightarrow X + Z \in L_k^2(\mathcal{S})$.
- By Theorem 5.3 in [7], $X \in M_k(\mathcal{S}), a \in \mathbb{R} \Rightarrow aX \in M_k(\mathcal{S})$, and $\mathbb{E}_P[\|X\|_2^2] < +\infty$ means $\mathbb{E}_P[\|aX\|_2^2] = a^2\mathbb{E}_P[\|X\|_2^2] < +\infty$. Therefore $X \in L_k^2(\mathcal{S}), a \in \mathbb{R} \Rightarrow aX \in L_k^2(\mathcal{S})$.

Based on the above foundations, now we prove the result in this Lemma. First suppose $Y \notin L_k^2(\mathcal{S})$, then

$$\mathbb{E}_P[\|Y\|_2^2] = \sum_{i=1}^{k} \mathbb{E}_P[Y_i^2] = +\infty\,, \tag{102}$$

thus there must be some $\mathbb{E}_P[Y_i^2] = +\infty$. Without loss of generality, suppose $\mathbb{E}_P[Y_1^2] = +\infty$. Since $\mathbb{E}_P[\|W\|_2^2] < +\infty \Rightarrow \mathbb{E}_P[W_1^2] < +\infty$, if $\mathbb{E}_P[(Y_1 - W_1)^2] < +\infty$ then

$$\begin{aligned}
&\mathbb{E}_P[Y_1^2] \\
&= \mathbb{E}_P[(Y_1 - W_1)^2] + \mathbb{E}_P[W_1^2] + 2\mathbb{E}_P[W_1(Y_1 - W_1)] \\
&\le \mathbb{E}_P[(Y_1 - W_1)^2] + \mathbb{E}_P[W_1^2] + \\
&\quad 2\sqrt{\mathbb{E}_P[W_1^2]\mathbb{E}_P[(Y_1 - W_1)^2]} \\
&< +\infty, && (103)
\end{aligned}$$

which is a contradiction. Therefore $\mathbb{E}_P[(Y_1 - W_1)^2] = +\infty$, which indicates

$$\mathbb{E}_P[\|W - Y\|_2^2] = +\infty,\ \forall\, Y \in M_k(\mathcal{S}) - L_k^2(\mathcal{S})\,. \tag{104}$$

For $Y \in L_k^2(\mathcal{S})$, We can write $Y = (Y_1, Y_2, \ldots, Y_k)$ and each Y_i is measurable from $(\Omega, \mathcal{S})$ to $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ by Lemma 7. And $Y \in L_k^2(\mathcal{S}) \Rightarrow \mathbb{E}_P[|Y_i|^2] < +\infty$. By Jensen inequality we have $\mathbb{E}_P[|Y_i|]^2 \le \mathbb{E}_P[|Y_i|^2] < +\infty$, therefore $\mathbb{E}_P[|Y_i|] < +\infty$, $\forall\, 1 \le i \le k$. And by Cauchy-Schwartz inequality $\mathbb{E}_P[|W_iY_i|] < +\infty$, $\forall\, 1 \le i \le k$. Then by Item 5 of Theorem 14.4 in [2], we have

$$\mathbb{E}_P[W_iY_i|\mathcal{S}] = Y_i\mathbb{E}_P[W_i|\mathcal{S}], \forall\, 1 \le i \le k, \forall\, Y \in L_k^2(\mathcal{S}). \tag{105}$$

Therefore we observe that $\forall\, Y \in L_k^2(\mathcal{S})$:

$$\begin{aligned}
&[W - \mathbb{E}_P[W|\mathcal{S}], Y] \\
&= [W, Y] - [\mathbb{E}_P[W|\mathcal{S}], Y] && (106)\\
&= \sum_{i=1}^{k} \mathbb{E}_P[W_iY_i] - \sum_{i=1}^{k} \mathbb{E}_P[\mathbb{E}_P[W_i|\mathcal{S}]Y_i] && (107)\\
&= \sum_{i=1}^{k} \mathbb{E}_P[W_iY_i] - \sum_{i=1}^{k} \mathbb{E}_P[\mathbb{E}_P[W_iY_i|\mathcal{S}]] && (108)\\
&= 0 && (109)
\end{aligned}$$

where (106) holds since $[.,.]$ is an inner product, (107) holds by (96) and Definition 2, and (108) holds by (105). (109) shows that

$$W - \mathbb{E}_P[W|\mathcal{S}] \perp L_k^2(\mathcal{S}). \tag{110}$$

By Definition 2 and the condition $\mathbb{E}_P[\|W\|_2^2] < +\infty$, $\mathbb{E}_P[W|\mathcal{S}] \in L_k^2(\mathcal{S})$. Now suppose $Z \in M_k(\mathcal{S})$ s.t.

$$\|W - Z\|^2 = \mathbb{E}_P[\|W - Z\|_2^2] = \inf_{Y \in M_k(\mathcal{S})} \mathbb{E}_P[\|W - Y\|_2^2], \tag{111}$$

then by (104), $Z \in L_k^2(\mathcal{S})$. Also, (97) and (111) show $\|W - Z\| = \inf_{Y \in L_k^2(\mathcal{S})} \|W - Y\|$. Hence by Lemma 10, $W - Z \perp L_k^2(\mathcal{S})$. Thus by (110), using Lemma 11, we have $Z = \mathbb{E}_P[W|\mathcal{S}]\ a.s.$ And it means

$$\arg\min_{Y \in M_k(\mathcal{S})}\ \mathbb{E}_P[\|W - Y\|_2^2] = \mathbb{E}[W|\mathcal{S}]\ a.s. \tag{112}$$

■

Proof of Theorem 4 (Continued) : We now resume the proof of Theorem 4. As defined earlier, $W^{(t)}$ is a square-integrable random vector representing the information of interest at time t, hence $\mathbb{E}[\|W^{(t)}\|_2^2] < +\infty$. And All Euclidean spaces below are equipped with their standard Borel σ-algebras, which we do not explicitly mention thereafter. Since we prove the general case, we need to redefine the two classes of functions and their theoretical discrepancy for $\forall\, k \in \mathbb{N}_+$:

$$\mathcal{F}_{\text{img},k} = \{f : \mathbb{R}^d \to \mathbb{R}^k \mid f \text{ measurable}\}; \tag{113}$$

$$\mathcal{F}_{\mathrm{vid},k}^{(l,r)} = \{f : \mathbb{R}^{(r+l+1)d} \to \mathbb{R}^k \mid f \text{ measurable}\}; \quad (114)$$

$$\Delta_{l,r}^k = \mathbb{E}\Bigg[\Bigg\| \arg\min_{f\in\mathcal{F}_{\mathrm{img},k}} \mathbb{E}\big[\|f(\mathcal{I}_t) - W^{(t)}\|_2^2\big] - \arg\min_{f\in\mathcal{F}_{\mathrm{vid},k}^{(l,r)}} \mathbb{E}\big[\|f(\mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}) - W^{(t)}\|_2^2\big]\Bigg\|_2^2\Bigg]. \quad (115)$$

Suppose the underlying sample space for the video is Ω, each $w \in \Omega$ is a complete elementary outcome generating a video. Now, by Lemma 9 we have

$$\{f(\mathcal{I}_t) : f \in \mathcal{F}_{\mathrm{img},k}\} = M(\sigma(\mathcal{I}_t)); \quad (116)$$

$$\{f(\mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}) : f \in \mathcal{F}_{\mathrm{img},k}^{(l,r)}\} = M(\sigma(\mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r})) \quad (117)$$

where the notation (93) in Definition 5 is used, then Lemma 12 and Definition 3 tell that

$$\begin{aligned}
&\arg\min_{f\in\mathcal{F}_{\mathrm{img},k}} \mathbb{E}\big[\|f(\mathcal{I}_t) - W^{(t)}\|_2^2\big] \\
&= \arg\min_{Y\in M(\sigma(\mathcal{I}_t))} \mathbb{E}\big[\|Y - W^{(t)}\|_2^2\big] \\
&= \mathbb{E}[W^{(t)} \mid \mathcal{I}_t] \; a.s.; \qquad (118)\\
&\arg\min_{f\in\mathcal{F}_{\mathrm{vid},k}^{(l,r)}} \mathbb{E}\big[\|f(\mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}) - W^{(t)}\|_2^2\big] \\
&= \arg\min_{Y\in M(\sigma(\mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}))} \mathbb{E}\big[\|Y - W^{(t)}\|_2^2\big] \\
&= \mathbb{E}[W^{(t)} \mid \mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}] \; a.s. \qquad (119)
\end{aligned}$$

Replacing them in (115), we proved

$$\Delta_{l,r}^k = \mathbb{E}\Big[\Big\|\mathbb{E}[W^{(t)} \mid \mathcal{I}_t] - \mathbb{E}[W^{(t)} \mid \mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}]\Big\|_2^2\Big]. \quad (120)$$

If $\mathbb{E}[W^{(t)} \mid \mathcal{I}_t] = \mathbb{E}[W^{(t)} \mid \mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}]$, then it's trivial that $\Delta_{l,r}^k = 0$ by (120). If $\mathbb{E}[W^{(t)} \mid \mathcal{I}_t] = \mathbb{E}[W^{(t)} \mid \mathcal{V}]$, then

$$\begin{aligned}
&\mathbb{E}[W^{(t)} \mid \mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}] \\
&= \mathbb{E}[\mathbb{E}[W^{(t)} \mid \mathcal{V}] \mid \mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}] \\
&= \mathbb{E}[\mathbb{E}[W^{(t)} \mid \mathcal{I}_t] \mid \mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}] \\
&= \mathbb{E}[W^{(t)} \mid \mathcal{I}_t]. \qquad (121)
\end{aligned}$$

Therefore $\Delta_{l,r}^k = 0$ again by (120). If $\mathbb{E}[W^{(t)} \mid \mathcal{I}_t] = W^{(t)}$, then

$$\begin{aligned}
&\mathbb{E}[W^{(t)} \mid \mathcal{I}_t] \\
&= \mathbb{E}[\mathbb{E}[W^{(t)} \mid \mathcal{I}_t] \mid \mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}] \\
&= \mathbb{E}[W^{(t)} \mid \mathcal{I}_{t-l},\dots,\mathcal{I}_{t+r}]. \qquad (122)
\end{aligned}$$

Hence $\Delta_{l,r}^k = 0$ by (120).

Thus, Theorem 4 in the main text follows immediately as a special case of the more general result proved above, obtained by instantiating $\Delta_{l,r}^k$ with $k = 1$ and replacing $W^{(t)}$ by $\mathcal{C}_t$. ■

Remark The quantity $\Delta_{l,r}$ in Theorem 4 plays the same conceptual role as classical theoretical quantities in statistical learning—such as the generalization error, Bayes risk, or mutual information—that are well-defined functionals characterizing the structure of a prediction problem, but not directly computable from finite samples. Our use of the term "quantifiable" refers to the fact that $\Delta_{l,r}$ is mathematically well-defined, internally consistent, and comparable across settings, rather than suggesting that it can be numerically estimated in closed form.

References

[1] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *Bmvc*, page 3, 2012. 1

[2] Bruce K Driver. Probability tools with examples. *Lectures notes available on B. Driver's page*, 2010. 10, 13

[3] Yanyan Fang, Biyun Zhan, Wandi Cai, Shenghua Gao, and Bo Hu. Locality-constrained spatial transformer network for video crowd counting. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 814–819. IEEE, 2019. 1

[4] Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer, 1998. 7

[5] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Estimating people flows to better count them in crowded scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 723–740. Springer, 2020. 1, 2

[6] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19618–19627, 2022. 4

[7] Samuel James Taylor. *Introduction to measure and integration*. CUP Archive, 1973. 9, 10, 11, 13

[8] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7812–7821, 2021. 1, 3