ShadowFlow: Learning Ambient Shadow Motion as a Non-Visual State Modality for Embodied Language Interaction

Anonymous ACL submission

Abstract

Language grounded embodied agents require accurate and continuous human state localization in indoor environments, but camera based tracking is often unacceptable in privacy sensitive applications. Existing device free approaches under unmodulated light lack a structured motion representation that can support sparse sensing and multi view sequence learning. To address this gap, we present ShadowFlow, a non imaging framework that infers continuous 2D trajectories from ambient illumination using sparse photodiode (PD) arrays without active modulation or visual capture. ShadowFlow lifts sparse PD readings into a differentiable grayscale shadow field on a virtual wall and derives a compact shadow flow tensor using lightweight optical flow operators. Since shadow deformation is view dependent and spatially heterogeneous, ShadowFlow encodes each view with attention parallel encoders and performs recurrent fusion to aggregate complementary spatial cues for trajectory regression. On 927 minutes of real world recordings from seven participants in two indoor layouts, ShadowFlow achieves centimeter level accuracy with a 2.35 cm mean localization error and supports real time inference on embedded hardware. The results indicate that ambient shadow flow provides a practical non visual motion modality that supports cross modal grounding for embodied language interaction and robotic perception.

1 Introduction

Large language models and language conditioned embodied agents are broadening the capabilities of interactive AI systems, spanning assistive robotics and intelligent physical spaces (Zhang et al., 2025a). Modern embodied agents depend on high quality perceptual inputs to ground decisions into actions, often relying on dense visual or proprioceptive streams, including egocentric video and motion sensors, to reason about the surrounding world (Wang et al., 2025; Zhang et al., 2025b). Despite their effectiveness, camera based human tracking is frequently restricted in real deployments, especially in private or human sensitive environments such as homes, workplaces, and elder care facilities, due to privacy regulations, user discomfort, calibration complexity, and sustained on device computation costs (Qiao et al., 2025; Cao et al., 2025). This motivates research on non visual sensing modalities that can supply reliable spatial state inputs without capturing identifiable visual content (Wu et al., 2024; Xiao et al., 2024).

Ambient visible light offers an attractive physical signal for passive state sensing because illumination is ubiquitous and already optimized for human comfort. When a moving person blocks and reroutes incident light rays, the resulting shadow deformation alters local irradiance, which can be measured by low cost photodiodes without forming images (Alijani et al., 2025). Prior visible light sensing studies demonstrate that shadows can support device free perception tasks, including posture inference, activity recognition, and coarse positioning (Zhang et al., 2025c; Gao et al., 2025). Recent neural architectures further exploit shadow features for localization, yet they typically assume dense sensing layouts or controlled emitters and reduce spatial motion to low dimensional proxies such as occupancy or region level labels, limiting their ability to decode continuous trajectories under naturally drifting illumination (Chen et al., 2025; Zhang et al., 2024). Separately, ambient light positioning systems that rely on commodity sensors focus on device based localization and depend on strong priors about calibrated beacons or user worn hardware, which shifts the problem away from purely environment grounded motion learning (Hegde et al., 2024; Zhang et al., 2025d). A key open question remains: whether sparse, non imaging photodiode arrays under unmodulated ambient illumination can produce a structured motion signal that supports continuous

and real time trajectory regression.

We argue that a structured motion representation is required to convert sparse shadow measurements into a learnable sequence for embodied state estimation. Optical flow in vision systems exemplifies this principle by transforming temporal brightness gradients into spatially coherent motion tensors. ShadowFlow adapts this intuition to the non visual regime by reconstructing a differentiable, grid aligned grayscale shadow field on a virtual wall from sparse photodiode readings and computing a compact shadow flow tensor that captures geometry conditioned brightness gradients. This intermediate representation enables motion to be expressed in relative brightness changes rather than absolute intensity, which simplifies global drift compensation, preserves spatial gradients for sequence learning, and supports interpretability through direct inspection of the reconstructed shadow structure.

On top of this representation, we propose ShadowFlow, a multi view spatiotemporal fusion framework that encodes each view with attention parallel encoders to preserve complementary spatial gradients, performs recurrent fusion to aggregate viewpoint specific shadow cues, and decodes continuous 2D trajectories using a temporal regression head. We validate ShadowFlow on a real world prototype deployed with multiple photodiode arrays in two indoor layouts under naturally varying, unmodulated illumination. Our long duration dataset includes 927 minutes of recordings from seven participants, and the evaluation demonstrates centimeter level localization accuracy with a 2.35 cm mean error while sustaining real time embedded inference. These results confirm that ambient shadow motion is a feasible non visual state modality for multimodal grounding in embodied AI and can serve as a reliable spatial state input for human centric interactive systems.

The contributions of this work are summarized as follows:

- We formalize ambient shadow motion as a non imaging optical flow representation by lifting sparse photodiode readings into a differentiable grayscale shadow field and its derived motion tensor for continuous localization.
- We introduce ShadowFlow, a multi view attention parallel recurrent architecture that fuses shadow flow features across photodiode arrays without cameras, tags, or active light modulation.

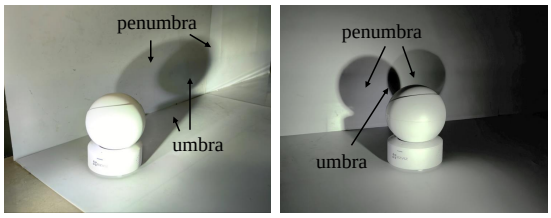


Figure 1: Penumbra and umbra in natural light.

- We build and evaluate a real world prototype and long duration dataset, demonstrating centimeter level accuracy with real time embedded trajectory decoding, which provides evidence for passive, privacy preserving embodied state sensing through ambient light shadows.

2 Theory of Light and Shadow Dynamics

ShadowFlow builds on basic principles of indoor illumination and shadow formation. This section summarizes how ambient light, human bodies, and walls interact to produce umbra and penumbra patterns, and how these patterns can be interpreted as a low resolution optical flow field observable by PD arrays. The qualitative behavior is illustrated in Fig. 1, which shows the inner umbra and outer penumbra regions under both natural and artificial light sources.

2.1 Indoor Illumination and Reflection

In our scenarios, indoor illumination comes from a combination of daylight through windows and ceiling lamps. We model each source as an extended emitter that generates a bundle of rays. When a ray with direction vector $\mathbf{i}$ hits a surface with unit normal $\mathbf{n}$, the specular reflection direction can be described by the standard reflection law

$$\mathbf{r} = \mathbf{i} - 2(\mathbf{i} \cdot \mathbf{n})\,\mathbf{n}, \tag{1}$$

where $\mathbf{r}$ is the reflected direction.

Indoor walls are predominantly Lambertian, scattering incident light across broad directions rather than generating a single specular reflection. When an object occludes part of the incoming light bundle, it casts a shadow on surrounding surfaces, and the intensity measured by a photodiode corresponds to the irradiance integrated over its field of view(FOV). As the subject moves, the shape of the occluded region varies, which subsequently modulates the irradiance received by each PD.

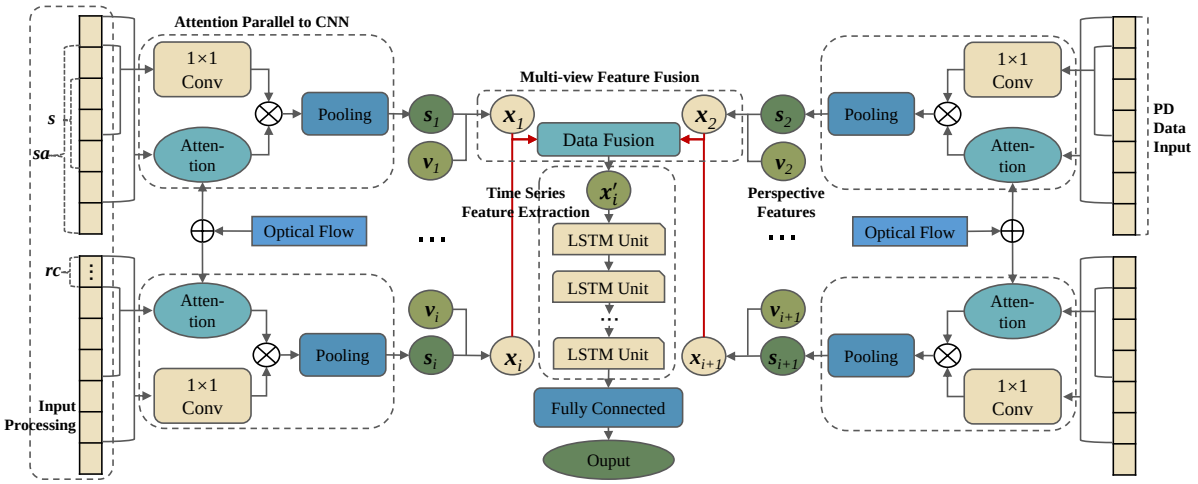


Figure 2: ShadowFlow pipeline. For each view, Input Processing extracts windows of length s and s_a (optionally using reference channel r_c). A 1×1 Conv branch and an attention branch are gated and pooled to produce embeddings (s_i, v_i), concatenated as x_i. Multi-view Data Fusion aggregates $\{x_i\}$ into x'_i, decoded by stacked LSTM units and a fully connected head to output the 2D position estimate.

2.2 Umbra, Penumbra, and Shadow Induced Intensity

The shadow cast by a human body can be decomposed into an umbra and a penumbra region. The umbra is the region where the light source is fully occluded, leading to a significant drop in irradiance. The penumbra corresponds to partial occlusion, where only part of the source is blocked and the intensity is reduced but non-zero. Fig. 1 highlight these regions under natural light and spot illumination, with a dark inner core (umbra) and a softer halo (penumbra).

Let $I(x, y, t)$ denote the brightness of a point on the wall at spatial coordinates (x, y) and time t. Conceptually, we write:

$$I(x, y, t) = I_0(x, y, t)\,\big(1 - \rho(x, y, t)\big), \quad (2)$$

where $I_0(x, y, t)$ is the background illumination in the absence of a subject, and $\rho(x, y, t) \in [0, 1]$ is an occlusion factor that increases from 0 (no shadow) to 1 (full umbra). When a person moves through the scene, the boundaries between different values of ρ sweep across the wall and create smooth spatiotemporal variations in $I(x, y, t)$.

Each PD i positioned at (x_i, y_i) measures the irradiance integrated over its FOV, which we approximate by sampling the brightness field at its center:

$$s_i(t) \approx I(x_i, y_i, t). \quad (3)$$

These samples are the basic inputs of ShadowFlow.

2.3 From Shadow Dynamics to Shadow Flow

To relate intensity changes to motion, we adopt the classical brightness constancy assumption along the motion of iso irradiance curves:

$$I_x\, u + I_y\, v + I_t \approx 0, \quad (4)$$

where I_x, I_y, and I_t are the spatial and temporal derivatives of brightness, and (u, v) is the optical flow vector describing the apparent motion of shadow patterns.

In our setting, we do not observe dense images but only the discrete PD samples $\{s_i(t)\}$. To obtain a continuous brightness field from discrete PD samples, we apply standard bilinear interpolation:

$$I(x, y, t) = \mathbf{w}(x, y)^\top \begin{bmatrix} I_{11}(t) \\ I_{21}(t) \\ I_{12}(t) \\ I_{22}(t) \end{bmatrix}, \quad (5)$$

where $\mathbf{w}(x, y)$ contains the four bilinear interpolation weights defined by the normalized offsets $(\Delta x, \Delta y)$.

We reconstruct a low resolution virtual wall shadow map $I(x, y, t)$, where each cell represents shadow induced brightness. The temporal dynamics of umbra and penumbra edges form the shadow flow signal that supports the multi view neural design in Section 3.

3 Method

ShadowFlow aims to recover fine grained human motion from sparse ambient light variations cap-

tured by multiple PD arrays. Building on the shadow formation model in Section 2, we treat shadow induced irradiance changes as a low resolution optical flow like field that evolves across walls and PD panels. As summarized in Fig. 2, ShadowFlow performs (i) input processing on raw PD time series, (ii) shadow image reconstruction and shadow flow estimation, (iii) attention parallel view encoding that produces per view embeddings (s_i, v_i), (iv) multi view data fusion to obtain x'_i, and (v) time series feature extraction using stacked LSTM units followed by a fully connected head to infer the 2D trajectory.

Let $\mathbf{p}_t = (x_t, y_t)$ denote the ground truth 2D position at time t. Suppose we have M views (e.g., wall panels). For each view $m \in \{1, \dots, M\}$, let $\mathbf{x}_t^{(m)} \in \mathbb{R}^{K_m}$ be the raw intensity vector from K_m PD elements at time t. ShadowFlow learns a mapping

$$f_\theta : \{\mathbf{x}_{t-L+1:t}^{(m)}\}_{m=1}^{M} \mapsto \hat{\mathbf{p}}_t, \tag{6}$$

where L is the temporal window length and θ denotes all learnable parameters. In the pipeline figure, the index i corresponds to the current prediction time step (i.e., $i \leftrightarrow t$).

3.1 Input Processing: Baseline Removal, Normalization, and Cross Scale Slicing

ShadowFlow operates on non-imaging PD time series that may exhibit slow drift due to daylight changes and lamp fluctuations. For each view m, we first remove global illumination trends using a baseline estimate:

$$\mathbf{y}_t^{(m)} = \mathbf{x}_t^{(m)} - \mathbf{b}_t^{(m)}, \tag{7}$$

where $\mathbf{b}_t^{(m)}$ is obtained via a sliding temporal median. When available, we additionally use a reference channel $r_c(t)$ (a PD minimally affected by human shadows) to stabilize the baseline under strong illumination changes, consistent with the r_c input in Fig. 2.

We then standardize each channel to reduce inter sensor bias:

$$\tilde{\mathbf{y}}_t^{(m)} = \frac{\mathbf{y}_t^{(m)} - \boldsymbol{\mu}^{(m)}}{\boldsymbol{\sigma}^{(m)} + \epsilon}, \tag{8}$$

where $\boldsymbol{\mu}^{(m)}$ and $\boldsymbol{\sigma}^{(m)}$ are computed over the training set and ϵ is a small constant.

Finally, for every prediction time t, we extract two overlapping temporal windows:

- an original scale window of length s for view encoding,
- a cross scale window of length $s_a > s$ for attention conditioning.

This cross scale slicing provides both local motion details and longer temporal context while preserving efficiency.

3.2 Shadow Image Reconstruction & Shadow Flow Estimation

PD arrays are non imaging sensors and do not directly expose spatial gradients. To enable optical flow reasoning, we reconstruct a low resolution shadow image on a regular grid for each view and then estimate a coarse shadow flow field.

Shadow image reconstruction. Let (x_i, y_i) denote the wall coordinates of PD i on view m. For any grid location (x, y) between four neighboring PDs, we approximate the intensity by applying the bilinear interpolation defined in Eq. (5):

$$I^{(m)}(x, y, t) \leftarrow \text{bilinear}\big(\{x_{t,i}^{(m)}\}\big), \tag{9}$$

which yields a grid form shadow image $I_t^{(m)}(u, v)$. We apply a light smoothing filter (e.g., a small Gaussian kernel) to suppress sensor noise and flicker while preserving shadow boundaries:

$$\tilde{I}_t^{(m)}(u, v) = \mathcal{S}\big(I_t^{(m)}(u, v)\big). \tag{10}$$

Shadow flow estimation. We estimate a coarse shadow flow field $\mathbf{F}_t^{(m)}(u, v) = (u^{(m)}, v^{(m)})$ by enforcing the brightness constancy constraint in Eq. (4) on the reconstructed shadow image $\tilde{I}_t^{(m)}(u, v)$ using finite difference gradient operators. This produces a low resolution flow that retains the dominant direction and relative magnitude of shadow motion without solving for dense pixel level correspondences.

Flow conditioning cue. Following Fig. 2, the optical flow module provides a compact flow cue for the attention branch. Concretely, we compute a lightweight descriptor

$$\mathbf{g}_t^{(m)} = g\Big(\mathbf{F}_t^{(m)}\Big), \tag{11}$$

e.g., via spatial pooling of flow magnitude and direction statistics, which is concatenated with cross scale temporal features to condition attention.

3.3 Attention Parallel View Encoding

For each view m, ShadowFlow encodes the reconstructed shadow image $\tilde{I}_t^{(m)}$ using the attention parallel to CNN block in Fig. 2, which contains a 1×1 Conv projection and a flow conditioned attention branch. The two branches are merged by element wise gating and pooling:

$$\mathbf{a}_t^{(m)} = \phi_{1\times1}(\tilde{I}_t^{(m)}), \tag{12}$$

$$\boldsymbol{\alpha}_t^{(m)} = \phi_{\text{att}}\big(\tilde{\mathbf{y}}_{t-s_a+1:t}^{(m)} \oplus g(\mathbf{F}_t^{(m)}) \oplus r_c(t)\big), \tag{13}$$

$$\mathbf{u}_t^{(m)} = \text{Pool}\big(\boldsymbol{\alpha}_t^{(m)} \odot \mathbf{a}_t^{(m)}\big), \tag{14}$$

where $\phi_{1\times1}$ is a single 1×1 convolution, $g(\mathbf{F}_t^{(m)})$ is the pooled shadow flow cue, $r_c(t)$ is the optional reference PD channel, $\odot$ is element wise gating, and $\text{Pool}(\cdot)$ denotes spatial pooling. The pooled feature $\mathbf{u}_t^{(m)}$ is linearly projected into appearance and motion embeddings $(s_i^{(m)}, v_i^{(m)})$ and concatenated to form the per view representation:

$$x_i^{(m)} = \big[\psi_s(\mathbf{u}_t^{(m)});\ \psi_v(\mathbf{u}_t^{(m)})\big], \tag{15}$$

which corresponds to the x_i node in Fig. 2. All views $\{x_i^{(m)}\}$ are then passed to the Data Fusion block to generate x_i' for trajectory decoding.

3.4 Temporal Fusion and Trajectory Regression

ShadowFlow fuses all per view embeddings $x_i^{(m)}$ $(i \leftrightarrow t)$ via the Data Fusion block in Fig. 2 to obtain a unified representation for temporal decoding. The fusion is a lightweight spatial aggregation:

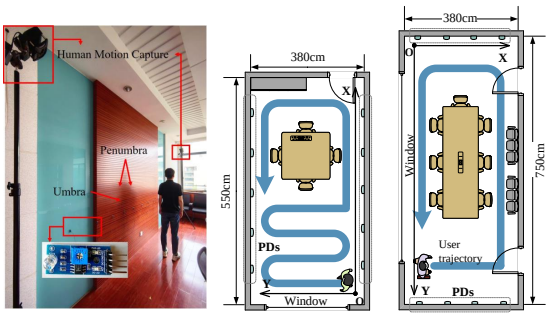
$$x_t' = \phi_{\text{fuse}}\Big(\big[x_t^{(1)}; \ldots; x_t^{(M)}\big]\Big), \tag{16}$$

where $\phi_{\text{fuse}}(\cdot)$ denotes a shallow stack of 1×1 convolutions followed by pooling. The fused feature x_t' is fed to a multi layer stacked LSTM to model shadow motion temporal evolution:

$$\mathbf{h}_t = \text{LSTM}(x_t', \mathbf{h}_{t-1}), \tag{17}$$

where the final layer hidden state $\mathbf{h}_t$ encodes the subject's trajectory history. A fully connected (FC) regression head estimates the 2D position:

$$\hat{\mathbf{p}}_t = W_{\text{fc}}\mathbf{h}_t + \mathbf{b}_{\text{fc}}, \tag{18}$$



(a) Experiment scenario in Scene_A (b) Scene_A: Office (c) Scene_B: Meeting-room

Figure 3: Experimental room structure and light source.

and parameters $\theta = \{W_{\text{fc}}, \mathbf{b}_{\text{fc}}, \phi_{\text{cnn}}, \phi_{\text{att}}, \phi_{\text{fuse}}\}$ are optimized by minimizing the mean squared trajectory error:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\mathbf{p}}_t \mathbf{p}_t^{\text{gt}}\|_2^2. \tag{19}$$

Due to the low dimensional PD input and minimal encoder decoder depth, ShadowFlow sustains sub millisecond inference latency, satisfying real time requirements for multimedia sensing and embodied interaction.

4 Experiments

This section evaluates the performance of ShadowFlow in terms of localization accuracy, robustness, generalization across environments, and computational efficiency. We first describe the dataset and experimental setup, followed by quantitative comparisons, ablation studies, and analysis of multi view fusion and temporal modeling. All experiments were conducted using the same hardware and evaluation protocol to ensure reproducibility.

4.1 Experimental Setup and Dataset Acquisition

ShadowFlow is assessed in two indoor environments (Fig. 3): Scene_A (Office, 380×550 cm) and Scene_B (Meeting Room, 380×750 cm). Three PD panels are installed on two walls orthogonal to the window plane at 1 m height, spaced at 30 cm intervals. A ceiling mounted PD provides the reference channel $r_c(t)$ to correct global illumination drift prior to shadow flow encoding. Each wall PD integrates irradiance within a 30° FOV over a 25 cm^2 active area, yielding sparse irradiance samples $x_{t,i}^{(m)}$. We collect 927 minutes of PD signals and motion capture 2D trajectories from 7 participants, partitioned into 70/15/15

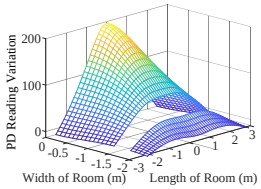


Figure 4: Motion consistent PD response over target trajectories.

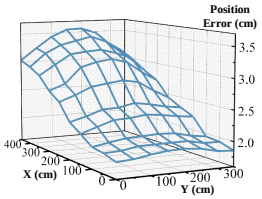


Figure 5: Spatial error distribution induced by shadow geometry.

for training/validation/testing. Preprocessed windows (s, s_a, r_c) are transformed into per view embeddings (s_i, v_i, x_i), fused as x'_i, and decoded by stacked LSTMs for 2D trajectory regression.

4.2 Baseline Methods and Evaluation Metrics

To contextualize performance, we compare ShadowFlow against cross modal and architectural baselines. Cross modal baselines include RF CSI tracking, thermal fusion sensing, and unmodulated visible light PD regression. Architectural baselines comprise an LSTM trajectory decoder, a recurrent model without view fusion, and an LSTM decoder paired with attention guided multi view fusion. Evaluation metrics cover MAE (cm), RMSE of trajectory deviation, DTW distance for temporal alignment, per frame inference latency, and memory footprint. We adopt sliding window preprocessing and report MAE as the main metric.

4.3 PD Reading Sensitivity to Target Location

We first verify a core assumption of ShadowFlow: sparse PD readings preserve spatial coherence under unmodulated illumination. We sample target locations on a 2D floor grid and record compensated responses from wall mounted PDs, applying ambient intensity normalization to suppress slow global brightness drift. Fig. 4 shows the resulting PD signal manifold over the floor plane. The surface is smooth with large scale gradients, indicating that PD responses are position conditioned rather than spatially flat or noisy. It also exhibits a dominant high response region and oblique attenuation, consistent with pose dependent occlusion of the incident light bundle.

4.4 Spatial Error Distribution

We next analyze spatial variation in localization accuracy. For each time step t, we compute the Euclidean error and aggregate errors by ground truth location. Fig. 5 visualizes the resulting spatial error field, which spans approximately 2.0 -3.5 cm in the example shown. The error surface exhibits structured heteroscedasticity rather than random scatter. Regions with stronger shadow observability yield lower error, while low contrast regions incur higher but bounded error, indicating graceful degradation. These spatially localized hard regions motivate ShadowFlow's attention weighted multi view fusion, which mitigates single view degeneracy and reduces view dependent ambiguity.

4.5 Participant Consistency

Fig. 6 illustrates that the normalized sparse PD signals produce consistent localization accuracy across participants. The median errors concentrate within a compact cm scale interval (approximately 1.6 to 2.8 cm) with small interquartile ranges, indicating that the learned shadow representation suppresses subject dependent variance before fusion. Participants with shorter whiskers exhibit reduced extreme outliers, suggesting that once shadow gradients are sufficiently discriminative, the model converges to a stable trajectory decoding regime without memorizing individual appearance. This supports that ShadowFlow learns view consistent, geometry aware shadow dynamics rather than overfitting to subject specific optical signatures.

4.6 Overall Accuracy vs Baselines

Fig. 7 compares the cumulative distribution of localization error between ShadowFlow and representative baselines. KNN with temporal interpolation, CNN only, LSTM only, and a convolutional recurrent fusion baseline accumulate probability mass more slowly, particularly in the low error region. ShadowFlow reaches about 0.7 CDF at 2 cm and 0.9 at 3 cm, showing faster error concentration and fewer high error outliers. The improvement persists in the tail, indicating that multi view shadow fusion reduces the frequency of large error deviations by preserving complementary spatial gradients that are lost in single view or framewise sequence models. These results validate that shadow flow as a non visual motion modality, combined with attention guided multi view temporal fusion, yields robust continuous localization for embodied multimodal systems.

4.7 Ablation Studies

Table 1 reports ablations that isolate the roles of shadow flow cues and multi view feature encoding. When the shadow flow branch is removed, errors increase across both scenes and across all statistics,

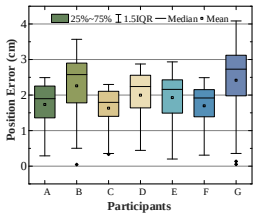


Figure 6: Localization error across different participants.

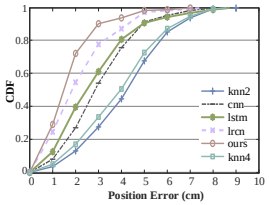


Figure 7: CDF of localization error comparing ShadowFlow.

Table 1: Localization position errors across scenes (cm).

Method	Scene	Error (cm) Mean	Med.	90th
Proposed and ablations				
CNN+ATT+LSTM	A	**2.35**	**2.33**	**2.60**
(w/ optical flow)	B	**2.49**	**2.47**	**3.01**
CNN+ATT+LSTM	A	3.34	3.31	3.63
(w/o optical flow)	B	3.87	3.82	4.24
ATT+LSTM	A	3.00	2.97	3.17
(w/ optical flow)	B	3.26	3.21	4.05
Baselines				
CNN+LSTM	A	7.92	7.85	8.13
	B	8.80	8.73	9.75
LRCN	A	4.56	4.48	5.02
(CNN+RNN fusion)	B	4.92	4.89	5.47
KNN	A	6.31	6.22	6.98
(fingerprint)	B	6.74	6.69	7.45
KNN +	A	5.18	5.12	5.86
temporal smoothing	B	5.43	5.39	6.11

which shows that the reconstructed shadow flow signal carries motion information that is difficult to recover from raw intensity sequences alone. In Scene A, the mean and 90th percentile errors rise from 2.35 cm and 2.60 cm to 3.34 cm and 3.63 cm, respectively. In Scene B, the corresponding values increase from 2.49 cm and 3.01 cm to 3.87 cm and 4.24 cm. The consistent degradation in the 90th percentile suggests that shadow flow features are particularly helpful for difficult cases where shadows are weak or view dependent.

We also ablate the view encoding stack by removing the CNN based encoder while retaining the shadow flow input. This variant (ATT+LSTM with optical flow) yields higher mean errors of 3.00 cm in Scene A and 3.26 cm in Scene B, compared with 2.35 cm and 2.49 cm for the full model. The effect is more pronounced in the tail, especially in Scene B where the 90th percentile increases from 3.01 cm to 4.05 cm. Taken together, these ablations support the design choice of treating ambient shadows as a non visual modality and learning multi

Table 2: Position errors under different light conditions.

Condition	Mean	Median	90th Perc.
Weather Conditions			
Sunny	2.25	2.33	2.60
Cloudy	2.71	2.69	2.84
Rainy	4.18	3.96	4.07
Time of Day			
Morning	3.01	2.93	3.12
Noon	2.16	2.19	2.25
Dusk	5.16	5.93	5.21

view fusion on top of a motion oriented intermediate representation.

4.8 Robustness to Illumination Conditions

Table 2 reports localization errors under diverse ambient light conditions. Performance is strongest under stable, high contrast illumination, with 2.25 cm error in sunny weather and 2.16 cm mean, 2.25 cm 90th percentile at noon, where shadow gradients are sharp and temporally consistent. Errors increase under low contrast or drifting light, reaching 4.18 cm in rainy conditions and 5.16 cm at dusk due to blurred shadow boundaries and amplified low frequency intensity drift. Importantly, the degradation is progressive rather than abrupt, showing that the model maintains temporal coherence by encoding motion in relative brightness changes instead of absolute intensity.

5 Conclusion

In this paper, we introduced ShadowFlow, a device free localization framework that reconstructs a differentiable shadow field from sparse photodiode measurements and captures motion geometry through a lightweight shadow flow representation. By fusing multi view sequences with attention based encoders and a recurrent trajectory regressor, ShadowFlow provides a practical nonvisual state modality for multimodal grounding in embodied language interaction and robotic perception while preserving privacy.

6 Limitations

ShadowFlow currently assumes that shadows maintain minimal geometric continuity across views. Extreme illumination instability, fast target motions beyond photodiode sampling rates, or highly

reflective dynamic surfaces may weaken shadow gradients and reduce trajectory fidelity. The current prototype evaluates human walking trajectories in two indoor layouts; broader environments and additional viewpoints require further validation. Future work will investigate adaptive fusion under more complex light paths and faster motion dynamics.

References

M. Alijani, C. De Cock, and W. Joseph. 2025. Device-free visible light sensing: A survey. *IEEE Communications Surveys Tutorials*.

Biwei Cao, Qihang Wu, Jiuxin Cao, Bo Liu, and Jie Gui. 2025. External reliable information-enhanced multimodal contrastive learning for fake news detection. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 31–39. AAAI Press.

Xiaocan Chen, Qilin Yin, Jiarui Liu, Wei Lu, Xiangyang Luo, and Jiantao Zhou. 2025. GLCF: A global-local multimodal coherence analysis framework for talking face generation detection. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 75–83. AAAI Press.

Fang Gao, Lei Shi, Jingfeng Tang, Jiabao Wang, Shaodong Li, Shengheng Ma, and Jun Yu. 2025. Visual and textual commonsense-enhanced layout learning for vision-and-language navigation. *IEEE Transactions on Automation Science and Engineering*.

Chaitra Hegde, Yashar Kiarashi, Amy D Rodriguez, Allan I Levey, Matthew Doiron, Hyeokhyen Kwon, and Gari D Clifford. 2024. Indoor group identification and localization using privacy-preserving edge computing distributed camera network. *IEEE journal of indoor and seamless positioning and navigation*, 2:51–60.

Haotian Qiao, Vidya Srinivas, Peter Dinda, and Robert Dick. 2025. Efficient video redaction at the edge: Human motion tracking for privacy protection. *ACM Transactions on Embedded Computing Systems*, 24(5s):1–22.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, and 1 others. 2025. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87.

Shirley Wu, Shiyu Zhao, Qian Huang, Kexin Huang, Michihiro Yasunaga, Kaidi Cao, Vassilis N. Ioannidis, Karthik Subbian, Jure Leskovec, and James Y. Zou. 2024. Avatar: Optimizing LLM agents for tool usage via contrastive reasoning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ziyang Xiao, Dongxiang Zhang, Xiongwei Han, Xiaojin Fu, Wing Yin Yu, Tao Zhong, Sai Wu, Yuan Wang, Jianwei Yin, and Gang Chen. 2024. Enhancing LLM reasoning via vision-augmented prompting. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jinfang Hu, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. 2025a. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 13032–13056. Association for Computational Linguistics.

Shao Zhang, Xihuai Wang, Wenhao Zhang, Chaoran Li, Junru Song, Tingyu Li, Lin Qiu, Xuezhi Cao, Xunliang Cai, Wen Yao, Weinan Zhang, Xinbing Wang, and Ying Wen. 2025b. Leveraging dual process theory in language agent framework for real-time simultaneous human-ai collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4081–4108. Association for Computational Linguistics.

Weichen Zhang, Chen Gao, Shiquan Yu, Ruiying Peng, Baining Zhao, Qian Zhang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025c. Citynavagent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 31292–31309. Association for Computational Linguistics.

X. Zhang, C. Li, and C. Wu. 2025d. Tapor: 3d hand pose reconstruction with fully passive thermal sensing for around-device interactions. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (UbiComp)*. ACM.