Self-Supervised Denoising of 4-Tap Indirect Time-of-Flight Measurements with Phasor Consistency

Saner Halil Baskaya, Stanford University, EE367

Abstract—Indirect Time-of-Flight (iToF) imaging estimates depth from the phase of modulated illumination. While modern sensors enable megapixel-scale depth imaging, measurements are heavily corrupted by photon shot noise, particularly at long distances or under low-SNR conditions. Classical denoising approaches typically operate on reconstructed depth maps or require clean ground-truth supervision, which is difficult to obtain in practice. In this work, we propose a self-supervised learning approach that directly denoises the raw 4-tap correlation measurements used in iToF sensing. Our method trains a four-channel residual U-Net using noisy measurement pairs in a Noise2Noise framework. To preserve the phase relationships required for accurate depth reconstruction, we introduce a phasor-consistency loss on the derived in-phase and quadrature components. The method is evaluated on 100-MP iToF captures of flat targets at distances between 1 and 5 meters. Experimental results show significant reductions in phase noise and depth variance, with depth RMS error improvements of up to 68%. These results demonstrate that self-supervised denoising at the measurement level can substantially improve depth precision without requiring clean ground-truth data. At inference time, the trained network denoises a single 4-tap measurement stack, after which standard phase reconstruction is applied.

Index Terms—High-resolution, LiDAR, denoising, Noise2Noise, 3D imaging, photon shot noise

1 Introduction

Indirect Time-of-Flight (iToF) cameras estimate depth by measuring the phase shift between emitted and received modulated illumination. These sensors have become widely used in applications including robotics, autonomous navigation, and computational imaging. Modern systems employ high-resolution pixel arrays capable of megapixel-scale depth sensing.

However, iToF measurements are fundamentally limited by photon shot noise and low signal-to-noise ratio (SNR), particularly at long ranges or under low illumination conditions. Noise in the correlation measurements propagates directly into phase estimation, which leads to unstable depth estimates and degraded precision.

Traditional approaches attempt to reduce noise through temporal averaging or spatial filtering. Temporal averaging improves SNR but reduces temporal resolution, while spatial filtering may blur depth boundaries and high-spatial-frequencies. More recently, supervised deep-learning approaches have been proposed for depth-map denoising. However, these methods require clean or close-to-clean ground-truth depth maps [1], [2], which are difficult to obtain in realistic imaging scenarios.

Self-supervised learning methods such as Noise2Noise [3], [4] have shown that denoising can be learned using only pairs of noisy observations. However, applying these approaches directly to depth maps may distort the underlying physical measurement structure of iToF signals.

In this work, we instead perform denoising directly on the raw 4-tap correlation measurements produced by the sensor. These measurements encode the phase of the modulated signal through a complex phasor representation. To ensure that denoising preserves phase information required for accurate depth reconstruction, we introduce a phasor-consistency loss that enforces consistency in the derived in-phase (I) and quadrature (Q) components.

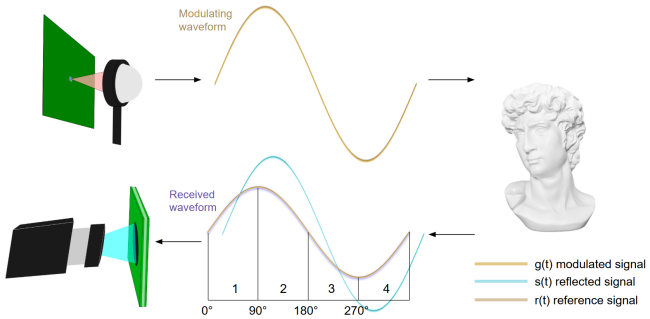

Fig. 1. Overview of the proposed iToF measurement pipeline

Instead of denoising reconstructed depth maps, we operate directly on the raw correlation measurements, allowing the model to suppress noise while preserving the physical structure required for phase-based depth estimation.

Our results demonstrate that physics-aware self-supervised denoising significantly improves phase stability and depth precision across a range of distances without requiring clean ground-truth supervision.

2 Related Work

Time-of-Flight Depth Denoising

Depth denoising in Time-of-Flight (ToF) imaging has been studied across both classical and learning-based paradigms. Traditional methods include temporal and spatial filtering techniques that improve signal stability at the

cost of temporal resolution or spatial detail. Recent work has leveraged convolutional neural networks to refine noisy depth maps, but these methods typically operate on reconstructed depth images rather than on the underlying raw correlation data, potentially altering the physical structure of the measurement and phase information.

A complementary set of approaches uses end-to-end learning frameworks that take raw ToF measurements as input and jointly perform tasks such as multi-path removal, phase unwrapping, and denoising. These methods generally require noisy-clean depth pairs during training, which are difficult to obtain and may suffer from misalignment or reconstruction errors when real ground truth is unavailable.

Self-Supervised and Unsupervised Learning

Self-supervised denoising techniques like Noise2Noise demonstrate that neural networks can learn to reduce noise using only pairs of independent noisy observations. Extensions such as Noise2Void remove the need even for paired noisy examples by learning from single noisy inputs with masking strategies. These methods have been applied primarily to natural image denoising and rely on statistical assumptions about noise. While self-supervised denoising techniques have been explored for raw depth image enhancement, handling the phasor-structured, signal-dependent noise in iToF systems remains challenging.

A recent work in time-of-flight imaging is the Self-Supervised End-to-End ToF Imaging framework based on RGB-D Cross-Modal Dependency [5], which proposes a self-supervised learning model that leverages the geometric information in RGB images as implicit supervision. By maximizing cross-modal mutual information between RGB and depth data, this method suppresses noise and preserves fidelity in ToF imaging without clean depth image pairs. A hybrid loss incorporating the statistical characteristics of raw measurements further improves robustness across multiple off-the-shelf cameras. This approach achieves performance competitive with supervised methods and generalizes across varying hardware settings, demonstrating the benefit of cross-modal cues in self-supervised ToF learning.

Positioning of Our Work

Our method differs from prior work in three key ways. First, rather than relying on cross-modal supervision from external RGB images, we focus on self-supervised learning solely from the raw 4-tap iToF measurements themselves, removing the need for additional modalities. Second, we introduce a phasor consistency loss that explicitly preserves the phase relationships used for depth estimation, addressing the risk that denoising might distort iToF phase information. Finally, by training directly on raw correlation measurements in a Noise2Noise framework, our model improves depth precision under realistic noise without requiring noisy-clean pairs, RGB guidance, or external depth supervision. Together, these design choices enable physics-aware denoising directly in the measurement domain while remaining fully compatible with standard iToF reconstruction pipelines.

3 METHOD

3.1 iToF Measurement Model

In a 4-tap indirect Time-of-Flight (iToF) system, each pixel records four correlation measurements corresponding to phase-shifted samples of the modulated optical signal:

$$x = (I_0, I_1, I_2, I_3). \tag{1}$$

These measurements encode the phase of the returning illumination signal, which is used to estimate scene depth. Following the standard demodulation model, the quadrature and in-phase components are computed as

$$Q = I_0 - I_2, \tag{2}$$

$$I = I_1 - I_3. \tag{3}$$

These define a complex phasor

$$z = I + jQ. \tag{4}$$

The phase of the received signal is obtained as

$$\phi = \operatorname{atan2}(Q, I), \tag{5}$$

and depth is proportional to the phase shift

$$d = \frac{c}{4\pi f_{\mathrm{mod}}}\phi, \tag{6}$$

where c is the speed of light and f_{mod} is the modulation frequency.

Because depth is derived from phase, noise in the raw tap measurements directly propagates into depth error. In particular, photon shot noise perturbs the tap differences, increasing phase variance and degrading depth precision, especially under low signal-to-noise ratio (SNR) conditions.

Moreover, the mapping from raw tap measurements to depth is highly nonlinear due to the atan2 phase computation and phase wrapping. Consequently, noise introduced at the measurement level becomes distorted after phase reconstruction, making post-hoc depth-map denoising less physically consistent than operating directly on the raw correlation measurements.

3.2 Self-Supervised Denoising Framework

To improve depth precision, we perform denoising directly on the raw 4-tap measurements prior to phase reconstruction. Instead of requiring clean supervision, we adopt a self-supervised Noise2Noise training strategy.

For a static scene, we assume access to two independent noisy observations of the same latent signal:

$$x = s + n_x, \qquad y = s + n_y, \tag{7}$$

where s denotes the unknown clean 4-tap measurement and n_x, n_y represent independent noise realizations.

A neural network f_θ is trained to map one noisy observation to another

$$\tilde{x} = f_\theta(x). \tag{8}$$

Under the Noise2Noise assumption of zero-mean independent noise, minimizing the mean squared error between $\tilde{x}$ and y encourages recovery of the underlying clean signal s without requiring ground-truth supervision.

Unlike generic image denoising, however, iToF measurements contain structured phase relationships across the four taps. Therefore, denoising must preserve the relative differences between taps to avoid distorting the reconstructed phase.

3.3 Network Architecture

We use a four-channel residual U-Net that takes the raw tap measurements (I_0, I_1, I_2, I_3) as input and predicts denoised taps of the same dimensionality. Joint processing of all four channels allows the network to exploit inter-tap correlations while preserving the physical structure of the iToF signal.

The network follows a standard encoder–decoder architecture with skip connections. Residual learning is employed so that the model predicts a correction term added to the input measurements, which stabilizes optimization and encourages the network to learn noise suppression rather than reconstructing the signal from scratch.

Training is performed on randomly sampled 96×96 patches extracted from independent captures of the same static scene.

3.4 Phasor Consistency Loss

A standard image reconstruction loss alone does not guarantee preservation of the phase relationships required for depth estimation. Small perturbations in tap differences may introduce bias in the recovered phasor.

To address this, we introduce a phasor consistency loss that constrains the derived in-phase and quadrature components.

From the predicted taps $\tilde{x} = (\tilde{I}_0, \tilde{I}_1, \tilde{I}_2, \tilde{I}_3)$ we compute

$$\tilde{Q} = \tilde{I}_0 - \tilde{I}_2, \tag{9}$$

$$\tilde{I} = \tilde{I}_1 - \tilde{I}_3. \tag{10}$$

Similarly, from the noisy target

$$y = (I_0^y, I_1^y, I_2^y, I_3^y) \tag{11}$$

we compute

$$Q_y = I_0^y - I_2^y, \tag{12}$$

$$I_y = I_1^y - I_3^y. \tag{13}$$

The full training objective is

$$\mathcal{L} = \lambda_{\text{img}} \, \text{MSE}(\tilde{x}, y) + \lambda_{\text{phasor}} \left[\text{MSE}(\tilde{I}, I_y) + \text{MSE}(\tilde{Q}, Q_y) \right]. \tag{14}$$

The first term corresponds to the standard Noise2Noise reconstruction loss. The second term explicitly enforces consistency of the phasor components used for depth reconstruction.

Directly penalizing phase is undesirable because phase is wrapped and becomes unstable when amplitude is low. Supervising the I/Q components avoids discontinuities while preserving the underlying phasor geometry.

3.5 Training and Reconstruction Pipeline

The overall pipeline is illustrated as follows. Independent noisy 4-tap captures of the same static scene are used to form training pairs. During training, one noisy capture is passed through the network while the other serves as the target.

At inference time, the trained model denoises the four tap measurements. The resulting taps are then processed using the standard iToF reconstruction pipeline: FFT-based demodulation, phase extraction, and phase-to-depth conversion.

Because the captured images are extremely large (100 MP), inference is performed using overlapping tiled processing.

The network processes the four correlation taps jointly and predicts a denoised 4-tap stack. At inference time, the model is applied to overlapping tiles of the full-resolution image. For a 96×96 patch, the forward pass requires a single U-Net evaluation, making the approach compatible with standard iToF processing pipelines.

4 Experimental Setup

4.1 Dataset

Experiments were conducted using high-resolution iToF captures of static planar targets at distances of 1.0, 1.5, 2.0, 2.5, 3.5, and 5.0 meters. Each scene was captured multiple times to obtain independent noisy observations required for self-supervised training.

The sensor resolution is approximately 9200×11264 pixels ($\sim$100 MP). Additional qualitative experiments include thin structures such as millimeter-scale wires and rods to evaluate reconstruction of small features.

4.2 Implementation Details

Training uses randomly sampled 96×96 patch pairs from independent noisy captures. The model is trained for 24 epochs using the Adam optimizer with learning rate 10^{-4} and batch size 16.

Images are normalized to the $[0, 1]$ range prior to training. At inference time, full-resolution frames are denoised using overlapping tiled inference.

4.3 Evaluation Methodology

Depth is reconstructed from the denoised taps using FFT demodulation followed by phase extraction and phase-to-depth conversion.

To evaluate depth precision, we restrict analysis to a region of interest corresponding to the planar target. Pixels are further filtered using amplitude-based gating to remove extremely low-SNR measurements. A robust plane fitting procedure with iterative outlier rejection is then applied.

To ensure fair comparison, raw and denoised reconstructions are additionally evaluated on the intersection of their valid masks.

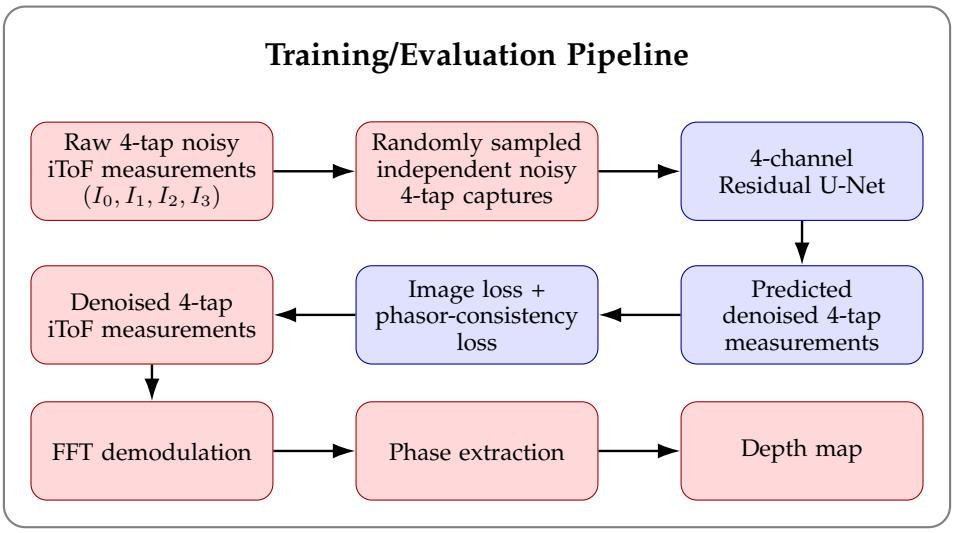

Fig. 2. Training and evaluation pipeline of the proposed method.

4.4 Evaluation Metrics

Performance is evaluated using two primary metrics:

1) **Phase standard deviation**, measuring temporal stability of the estimated phase.
2) **Plane-fit depth RMS**, computed from residuals after fitting a plane to the ROI.

These metrics quantify both signal-level stability and practical depth precision.

5 Results

5.1 Quantitative Results

We evaluate the proposed method on high-resolution iToF captures at distances ranging from 1 m to 5 m. Quantitative analysis focuses on planar regions of interest where the underlying geometry is approximately known. Depth precision is measured using two complementary metrics: phase standard deviation and plane-fit depth RMS within the selected regions of interest.

The primary comparison is against the standard iToF reconstruction pipeline applied directly to the raw 4-tap measurements, i.e., without learned denoising. This baseline is appropriate because most existing learning-based denoising methods operate on reconstructed depth maps or require auxiliary modalities such as RGB input, whereas our method operates directly on raw 4-tap correlation measurements prior to phase reconstruction. To ensure fair evaluation, we additionally compare raw and denoised reconstructions on the intersection of their valid masks.

Table 1 summarizes quantitative performance across target distances. The proposed method consistently reduces both phase variance and depth RMS relative to raw measurements.

The method achieves substantial improvements in depth precision across most distances. For example, at 1 m the plane-fit RMS decreases from 207.7 mm to 66.9 mm, corresponding to a 67.9% improvement. The trend across distances is visualized in Fig.4, which plots the plane-fit depth RMS as a function of range for both raw and denoised reconstructions.

TABLE 1
Quantitative results for flat targets at different distances.

d (m)	$\sigma_{\phi,\mathrm{raw}}$ (rad)	$\sigma_{\phi,\mathrm{den}}$ (rad)	$\Delta\phi$ (%)	$\mathrm{RMS}_{\mathrm{raw}}$ (mm)	$\mathrm{RMS}_{\mathrm{den}}$ (mm)	$\Delta\mathrm{RMS}$ (%)
1.0	0.0524	0.0305	41.2	207.7	66.9	67.9
1.5	0.0626	0.0286	54.0	244.2	89.5	64.0
2.0	0.0678	0.0278	59.5	254.2	96.2	62.2
2.5	0.0823	0.0360	55.8	266.5	158.9	40.3
3.5	0.1344	0.0419	68.2	325.2	185.2	42.8
5.0	0.0000*	0.0507	23.8*	277.9	209.9	18.1

* At 5 m, extremely low SNR makes amplitude gating unstable, so phase improvement should be interpreted cautiously.

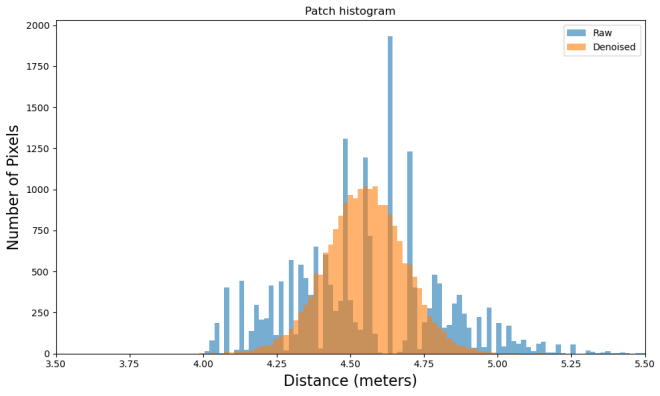

Fig. 3. Depth stability analysis within the diagonal region of interest. Top: raw and denoised depth maps within the ROI. Bottom: corresponding depth histograms. Denoising significantly reduces the spread of the depth distribution, indicating improved phase stability.

5.2 Qualitative Results

Qualitative examples further demonstrate the benefits of the proposed approach. For a diagonal plate captured at 5 m, the depth standard deviation within the region of interest decreases from approximately 25.6 cm to 13.5 cm after denoising. In addition, the denoised depth maps reveal small scene structures such as thin wires and rods that are difficult to observe in the raw reconstruction.

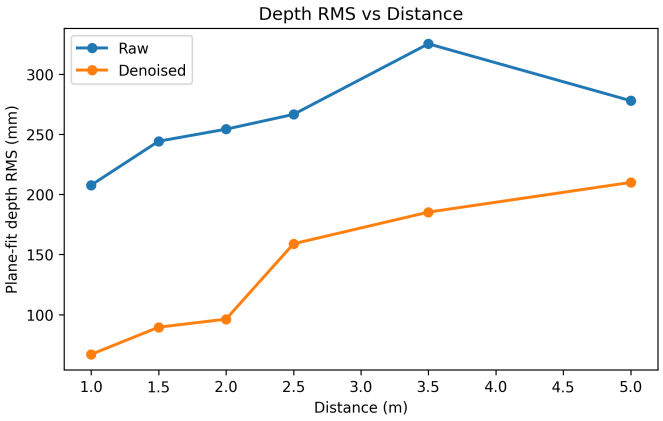

Fig. 4. Plane-fit depth RMS as a function of distance for raw and denoised reconstructions. Denoising improves depth precision across the evaluated range. Note that improvements at 5 m should be interpreted cautiously due to unstable low-SNR gating.

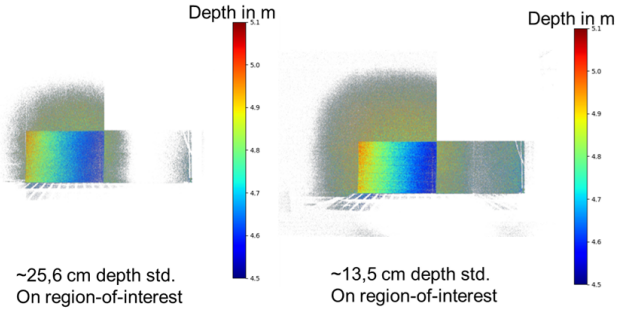

Fig. 5. Depth reconstruction of a diagonal plate at 5 m. Left: raw reconstruction showing strong phase noise. Right: reconstruction after denoising of the raw 4-tap measurements. The depth standard deviation within the region of interest decreases from approximately 25.6 cm to 13.5 cm, demonstrating improved phase stability and depth precision.

6 Discussion

These results demonstrate that self-supervised denoising applied directly to raw 4-tap measurements can significantly improve phase stability and depth precision. Because phase is computed from differences between correlation taps, small perturbations in raw measurements can significantly amplify phase estimation error. By denoising the taps directly while preserving the phasor geometry, the proposed method stabilizes phase estimation before depth reconstruction.

The proposed phasor consistency loss plays an important role by constraining the in-phase and quadrature components of the predicted taps. This prevents denoising from introducing phase bias while still allowing substantial noise suppression.

7 Limitations and Future Work

First, training requires independent noisy captures of the same static scene, which limits direct application in dynamic environments where multiple observations of identical geometry may not be available. Second, the current evaluation focuses primarily on planar targets, which simplifies

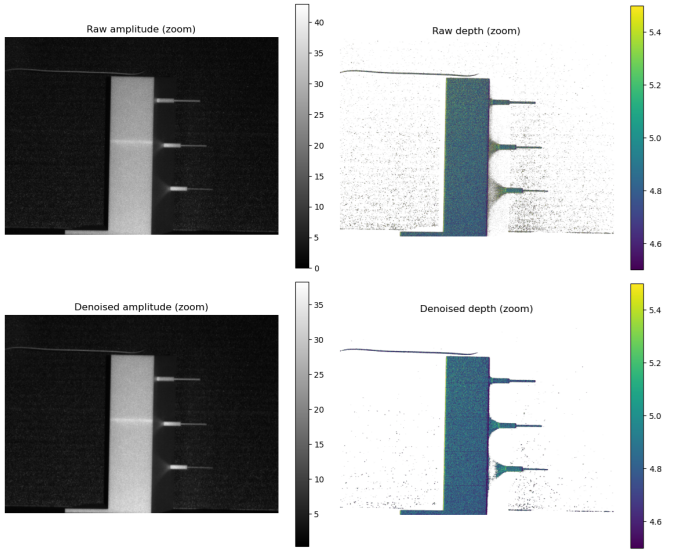

Fig. 6. Qualitative example showing reconstruction of fine structures. Top row: raw amplitude and raw depth reconstruction. Bottom row: denoised amplitude and depth after applying the proposed self-supervised model. Denoising suppresses shot noise while preserving thin structures such as wires and rods.

precision analysis but does not fully represent complex real-world scenes.

Future work could explore alternative self-supervised strategies that do not require paired observations, such as blind-spot networks or masked prediction. Incorporating explicit noise modeling or jointly optimizing denoising and depth reconstruction may further improve performance.

8 Conclusion

We presented a self-supervised denoising framework for 4-tap indirect Time-of-Flight measurements that improves phase stability and depth precision without requiring clean ground-truth supervision. By combining Noise2Noise training with a phasor consistency loss, the proposed approach suppresses measurement noise while preserving the phase relationships required for depth reconstruction.

Experiments on high-resolution iToF data demonstrate substantial improvements in both phase variance and plane-fit depth RMS across multiple distances. These results suggest that physics-aware denoising at the measurement level is a promising direction for improving high-resolution depth sensing systems. More broadly, these results highlight the potential of measurement-domain learning approaches for improving high-resolution depth sensing systems.

References

[1] Y. Dong, X. Zhang, and Z. Xiong, "Spatial hierarchy aware residual pyramid network for tof depth denoising," in *European Conference on Computer Vision (ECCV)*, 2020.

[2] J. Yan, Y. Wang, Y. Kao, M. Gong, J. Shen, and Y. Fu, "Ddrnet: Depth map denoising and refinement for consumer depth cameras," in *European Conference on Computer Vision (ECCV)*, 2018.

[3] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[4] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void: Learning denoising from single noisy images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[5] S. Wang, Y. Zhang, Y. Li, and L. Chen, "Self-supervised end-to-end time-of-flight imaging based on rgb-d cross-modal dependency," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.