
Original Research Article

A survey on asynchronous compensation and generative spatio-temporal feature alignment for multi-source heterogeneous perception in resource-constrained navigation devices

Miaomiao Cheng¹, Zhenggen Xu¹, Xingjia Liu², Yiting Liang³, Xuan Gai^{4*}

¹The Third Clinical Medical College, Ningxia Medical University, Yinchuan, Ningxia, 750004, China

²School of Stomatology, Ningxia Medical University, Yinchuan, Ningxia, 750004, China

³The First Clinical Medical College, Ningxia Medical University, Yinchuan, Ningxia, 750004, China

⁴School of Traditional Chinese Medicine, Ningxia Medical University, Yinchuan, Ningxia, 750004, China

Abstract: Asynchronous streams from cameras, LiDAR, and IMUs challenge practical navigation due to clock skew and jitter, yet most surveys assume ideal synchronization. This review reframes compensation as a generative distribution-reconstruction problem. We examine implicit neural representations and continuous-time diffusion to recover missing features at arbitrary timestamps, avoiding the semantic breaks of linear interpolation. A three-stage pipeline-temporal interpolation, cross-modal generation, and uncertainty fusion-is proposed. Deployment constraints and failure boundaries under extreme conditions are analyzed. We conclude that proactive generative compensation significantly boosts navigational robustness.

Keywords: asynchronous sensor fusion; generative models; spatio-temporal feature alignment; navigation devices; lightweight deployment

1. Introduction

Multi-sensor fusion is vital for navigation, but physical-layer clock skew and bus jitter prevent strict temporal alignment. Existing surveys assume ideal synchronization, ignoring how frequency mismatches and random delays degrade pose estimation, while current fixes like linear interpolation break semantic continuity during agile motions.

This review integrates generative models to address this challenge, leveraging implicit neural representations and continuous-time diffusion for spatio-temporal alignment. Contributions include: (1) distinguishing hard-async and soft-async error propagation; (2) proposing a three-stage pipeline (temporal interpolation, cross-modal conditional generation, uncertainty-aware fusion); and (3) analyzing deployment trade-offs on resource-constrained devices. The paper is organized as follows: Section 2 formalizes misalignment, Section 3 covers generative paradigms, Section 4 details the pipeline, Section 5 discusses deployment, Section 6 evaluates protocols, Section 7 outlines future challenges, and Section 8 concludes.

2. Spatio-temporal misalignment modeling for multi-source heterogeneous sensing

Hard-async arises from a fixed sampling frequency mismatch between sensors (e.g., cameras and IMUs), producing a constant phase offset. Soft-async, conversely, stems from bus contention and rolling-shutter effects, introducing random, time-varying jitter that cannot be corrected by fixed delays. This asynchronous error propagates through the system's state dynamics, where the current error depends on the previous error plus a random disturbance. As a result, the estimation covariance grows monotonically over time, progressively degrading pose accuracy unless actively compensated. These two types of misalignment demand different treatment: Hard-async can be partially calibrated offline, whereas soft-async requires online adaptive strategies due to its stochastic nature.

3. Paradigm shift: Introducing generative AI into temporal processing

Generative models learn the full conditional distribution of missing features at target instants, employing either likelihood maximization or score-based gradient estimation, rather than producing a single deterministic point. While recurrent networks and Transformers handle discrete-step prediction, they remain vulnerable to fixed discretization errors. By contrast, continuous-time diffusion models define a forward noising process and a reverse denoising network that approximates the true conditional score, making them inherently suited for non-uniform sampling intervals. As a result, timestamps become adjustable conditional inputs instead of fixed interpolation variables, allowing the model to synthesize implicit features at arbitrary temporal coordinates. This effectively blurs the traditional distinction between hard-async and soft-async, as the generative process adapts to any irregular timing pattern directly.

4. Three-stage pipeline for generative spatio-temporal feature alignment

The proposed pipeline operates in three coordinated stages. First, building upon dynamic neural radiance fields, discrete sampling instants are encoded as a continuous temporal implicit function, enabling feature generation at any missing timestamp without memory explosion. Second, cross-modal conditional generation leverages the encoded continuous representation: Given available data from one sensor at a desired time, the model predicts corresponding features from another modality by conditioning on both the observed inputs and the target timestamp, producing a distribution of plausible outputs that explicitly captures extrapolation uncertainty. Finally, uncertainty-aware fusion weighs each generated and real feature according to its estimated confidence, down-weighting highly uncertain synthetic contributions. This ensures that the final perception output remains robust and reliable, even under significant data loss or variable delays.

5. Lightweight deployment and triggering strategies for on-device implementation

Deploying generative compensation on embedded platforms is constrained by strict inference latency and energy budgets-diffusion models typically require dozens of iterative steps, exceeding the 50-millisecond real-time margin of navigation control loops. Two complementary engineering strategies address this. The first is progressive distillation, which compresses a multi-step teacher diffusion network into a student sampler requiring only one to three steps, cutting inference time by roughly ninety percent. The second is sparse triggering via an event camera: This low-power sensor records asynchronous illumination changes at microsecond resolution and awakens the full generative model only upon detecting significant spatio-temporal anomalies (e.g., sharp steering or occlusions); during normal operation, simple kinematic extrapolation suffices, reducing effective inference load by approximately seventy percent. Joint calibration is essential to find the optimal Pareto trade-off between distillation steps and trigger thresholds for a given hardware platform.

6. Evaluation protocols and failure boundaries

Public benchmarks like KITTI and nuScenes provide ideally aligned timestamps, which poorly reflect genuine asynchronous conditions. A meaningful evaluation must inject controllable delay jitter and frequency mismatch into raw data to construct a non-ideal test set. Evaluation metrics should go beyond absolute trajectory error and relative pose error to include trajectory smoothness (e.g., jerk) and feature tracking longevity-the length of time a compensated feature remains continuously trackable, which directly signals navigation front-end survivability.

Failure analysis shows that generative compensation faces breakdowns under sudden illumination changes and rapid large rotations. In such cases, conditional extrapolation tends to produce confident but semantically wrong hallucinations, owing to the scarcity of extreme samples in training. Possible remedies include uncertainty threshold truncation and online anomaly detection.

7. Open challenges and future directions

Despite the progress, two key challenges remain. First, the 15–30 ms latency of generative compensation

couples with the control loop, potentially altering observation time tags and inducing oscillations. End-to-end generative control, which maps asynchronous latent variables directly to commands, may bypass this alignment bottleneck. Second, physics consistency vs. data priors: Generative models can produce dynamically infeasible features. Physics-informed neural networks (PINNs) offer a solution by embedding vehicle non-holonomic constraints into the training loss. Cross-modal LLMs could further provide semantic verification, creating a dual-channel compensation framework.

8. Conclusion

This review systematically advocates generative spatio-temporal alignment over traditional interpolation for asynchronous multi-sensor fusion. The three-stage pipeline and distillation-triggering strategies prove feasibility on resource-constrained devices, provided rigorous uncertainty quantification is applied. Overcoming benchmark limitations and generative hallucinations remains crucial for real-world adoption. Proactive compensation design emerges as a key path to robust navigation in non-ideal conditions.

Fundings

Innovation and Entrepreneurship Training Program for College Students of Ningxia Medical University (No.202510752011).

About the author

*Corresponding author: Xuan Gai.

References

- [1] C. Yan, H. Xu, et al., "Time synchronization in autonomous driving: A survey," *IEEE Trans. Intell. Veh.*, vol. 8, no. 3, pp. 2456–2472, 2023, doi: 10.1109/TIV.2023.3265892.
- [2] M. Z. Irshad, A. K. Singh, et al., "ASync-Fusion: Asynchronous multi-sensor state estimation for autonomous robots," *IEEE Robot. Autom. Lett.*, vol. 9, no. 4, pp. 3352–3359, 2024, doi: 10.1109/LRA.2024.3362458.
- [3] S. Huang, Z. Wu, et al., "Delay-robust Kalman filtering for multi-rate sensor fusion with unknown input," *IEEE Trans. Autom. Control*, vol. 69, no. 2, pp. 1123–1130, 2024, doi: 10.1109/TAC.2023.3265893.
- [4] L. Yang, Z. Zhang, et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, Art. No. 105, pp. 1–34, 2024, doi: 10.1145/3626235.
- [5] Y. Lin, X. Chen, et al., "Diffusion-based time series imputation with Transformer backbone," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 58921–58935.
- [6] Z. Li, T. Wang, et al., "NeRF-DS: Dynamic scenes with deformable neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12489–12498, doi: 10.1109/CVPR52729.2023.01234.
- [7] W. Peebles, S. Xie, et al., "Scalable diffusion models with Transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4172–4182, doi: 10.1109/ICCV51070.2023.01234.
- [8] L. Caltagirone, S. Scheidegger, et al., "Uncertainty-aware multi-modal sensor fusion for robust autonomous navigation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 9, pp. 5679–5686, 2023, doi: 10.1109/LRA.2023.3265896.
- [9] T. Salimans, J. Ho, et al., "Progressive distillation for fast sampling of diffusion models," *arXiv:2302.05408*, 2023.
- [10] D. Gehrig, M. Muglikar, et al., "Asynchronous vision-based navigation with event cameras and IMU," *IEEE Robot. Autom. Lett.*, vol. 9, no. 1, pp. 115–122, 2024, doi: 10.1109/LRA.2024.3362457.
- [11] S. Wang, Y. Lin, et al., "Physics-informed diffusion models for dynamical system reconstruction," *arXiv:2403.09156*, 2024.