IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Game-KFS: Game-Theory-Inspired Keyframe Selection for Hybrid Representation Visual SLAM

Shilang Chen[1], Bo Yang[2], Chaoqun Wang[3], Peidong Fang[1], Haifei Zhu[1], Weinan Chen[4], and Yisheng Guan[1]

***Abstract*—Hybrid representation Visual Simultaneous Localization and Mapping (VSLAM) systems combine the inherent strengths of both discrete and field representations. They promise high-precision tracking and photo-realistic dense mapping. However, current keyframe selection methods in hybrid representation VSLAM struggle to satisfy both the high-precision tracking requirements of discrete representations and the high-quality rendering requirements of field representations. In this paper, we propose a game-theory-inspired keyframe selection approach that addresses the requirements of both representation types. We introduce two objective functions to comprehensively assess discrete point tracking and radiance field model rendering. By employing a game-theory-inspired framework, our method effectively balances these objectives to achieve improved keyframe selection. Experimental results demonstrate that integrating our approach into a hybrid representation VSLAM system significantly enhances tracking accuracy and rendering quality, outperforming existing keyframe selection methods.**

***Index Terms*—SLAM, Visual Tracking, Keyframe Selection, Hybrid Representation**

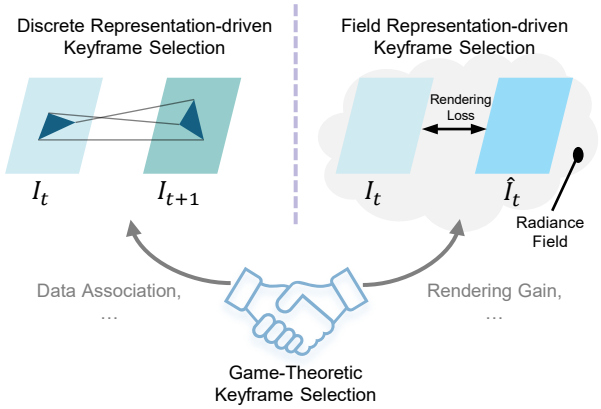

Fig. 1: Motivation for game-theory-inspired hybrid representation keyframe selection. Existing keyframe selection methods typically focus exclusively on either discrete geometric tracking or high-quality scene model rendering. Our game-theory-inspired approach balances these objectives by jointly optimizing high-precision geometric tracking and superior field model rendering.

I. Introduction

VSLAM [1]–[3] plays a pivotal role in autonomous mobile robots, unmanned aerial vehicles, and intelligent driving systems. Traditional VSLAM approaches based on discrete representations [4], [5] provide reliable localization but yield sparse, fragmented reconstructions, limiting their applicability in scenarios that demand high-fidelity, dense environmental models. In contrast, recent field representation-based VSLAM methods, leveraging Neural Radiance Fields (NeRF) [6] or 3D Gaussian Splatting (3D GS) [7]–[9], have demonstrated photo-realistic rendering and dense mapping capabilities. However, these field representation-based methods typically exhibit lower localization precision [10], owing to their fundamentally different representation optimization strategies.

To address these challenges, researchers have begun exploring hybrid representation VSLAM systems [11]–[16] that aim to leverage the complementary strengths of discrete feature tracking and field-based dense modeling. In such frameworks, keyframe selection [17] is critical: it must balance the computational demands and precision of discrete representations with the coverage and information requirements of field representations, directly impacting real-time performance, localization accuracy, and rendering fidelity.

Existing methods for keyframe selection in hybrid VSLAM can be divided into two categories. The first category relies solely on discrete representation tracking [11]–[13], prioritizing pose accuracy but ignoring the requirements of field-based rendering. The second category adopts a two-stage pipeline [14]–[16], where keyframes are first selected to satisfy discrete modeling needs and then refined for field reconstruction. However, this two-stage approach often prunes keyframes too aggressively in the initial phase, leaving insufficient information for subsequent field reconstruction and degrading the final rendering quality.

Motivated by this gap, we propose *Game-KFS*, a novel *Game*-theory-inspired *K*ey*F*rame *S*election method for hybrid representation VSLAM. As illustrated in Fig. 1, Game-KFS formulates keyframe selection as a strategic game between two players, one optimizing discrete geometric tracking and the other optimizing field-based rendering, thereby achieving

Manuscript received: June 23, 2025; Revised August 31, 2025; Accepted September 29, 2025.

This paper was recommended for publication by Editor J. Civera upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 62103179 and 52475008, in part by the Guangzhou Basic and Applied Basic Research Foundation under Grant 2024A04J4070, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010194. (Corresponding authors: Weinan Chen; Yisheng Guan.)

[1]Shilang Chen, Peidong Fang, Haifei Zhu, and Yisheng Guan are with the Biomimetic and Intelligent Robotics Lab, School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou, China. (e-mail: earcsirius@gmail.com, ysguan@gdut.edu.cn)

[2]Bo Yang is with the School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China.

[3]Chaoqun Wang is with the School of Control Science and Engineering, Shandong University, Jinan, China.

[4]Weinan Chen is with the State Key Laboratory of Precision Electronic Manufacturing Technology and Equipment, Guangdong University of Technology, Guangzhou, China. (e-mail: chenwn@gdut.edu.cn)

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

an optimal trade-off between high-precision localization and photo-realistic scene reconstruction. To the best of our knowledge, this is the first work to employ a game-theory-inspired framework for keyframe selection in VSLAM.

Our main contributions are summarized as follows:

- We propose a *game-theory-inspired, dynamically-weighted multi-objective* framework for online keyframe selection that balances the competing requirements of discrete (feature-based) tracking and field (radiance-field) rendering.
- We analyze the distinct requirements of discrete and field-based representations and derive objective functions that guide the keyframe selection process to meet both localization accuracy and rendering fidelity.
- We integrate the proposed Game-KFS strategy into a state-of-the-art hybrid representation VSLAM system. Extensive experiments on public benchmarks demonstrate that our method significantly outperforms existing approaches, achieving superior tracking accuracy and scene reconstruction quality.

II. Related Work

A. Hybrid Representation VSLAM

Traditional VSLAM systems achieve real-time, high-precision localization with discrete representations [10]. Methods such as ORB-SLAM3 [4] extract sparse feature points and perform geometric registration and bundle adjustment, while direct methods like DSO [5] optimize photometric error over image sequences to reconstruct a sparse map. Although these approaches excel at accurate localization, their sparse, fragmented reconstructions limit downstream tasks that require dense, continuous scene models.

With the advent of NeRF [18] and 3D GS [7], a new class of field representation-based VSLAM methods [10] has emerged. Approaches such as NICE-SLAM [6], SplaTAM [9], and MonoGS [8] jointly optimize scene radiance and density to produce photo-realistic, dense reconstructions. However, their tracking modules typically rely on rendering-based losses, which can lead to suboptimal localization accuracy.

Hybrid representation VSLAM systems [10], which synergize the strengths of discrete feature tracking and dense field-based mapping, have become a prominent direction in recent research. Methods such as Orbeez-SLAM [11], NGEL-SLAM [12], Photo-SLAM [13], NeRF-VO [14], MGS-SLAM [15], and MoD-SLAM [16] have demonstrated that a carefully designed hybrid framework can achieve both centimeter-level localization accuracy and photo-realistic dense reconstruction. These breakthroughs highlight the potential of hybrid VSLAM to overcome the sparse-map limitations of traditional discrete methods and the tracking challenges of pure field-based approaches.

B. Keyframe Selection in Hybrid Representation VSLAM

Keyframe selection is critical in VSLAM pipeline, as it determines the trade-off between computational efficiency, localization accuracy, and mapping quality [17]. In discrete representation-based systems, criteria for keyframe insertion include pose change thresholds, feature overlap, and temporal spacing, ensuring reliable tracking and accurate sparse map updates [4], [5]. Conversely, field representation-based systems select keyframes to maximize rendering quality or information gain in the learned scene representation, typically by evaluating photometric error, view-dependent uncertainty, or model convergence metrics [6], [8].

In hybrid representation VSLAM, existing keyframe selection strategies fall into two categories. The first directly adopts discrete tracker keyframes for field mapping (e.g., Orbeez-SLAM [11], NGEL-SLAM [12], Photo-SLAM [13]), thus ignoring the distinct requirements of dense reconstruction. The second performs a two-stage selection, first for discrete tracking and then for field mapping (e.g., NeRF-VO [14], MGS-SLAM [15], MoD-SLAM [16]). However, aggressive pruning in the initial stage often removes views that are crucial for high-quality rendering.

An effective approach must jointly consider the objectives of both representations. To address this gap, our work formulates keyframe selection in hybrid representation VSLAM as a sequential continuous optimization problem. This formulation explicitly balances the need for accurate discrete geometric tracking with the demand for high-quality radiance field model rendering, leading to a more comprehensive strategy.

C. Game-Theory-Inspired Perspectives for Keyframe Selection

In hybrid representation VSLAM frameworks, where both field and discrete representations are integrated, each keyframe influences two distinct modeling processes. Methods designed for field representations typically require images with diverse viewpoints and consistent appearances [8]. Such properties are essential for continuously optimizing rendering quality and achieving dense scene reconstruction. In contrast, keyframe selection for discrete representations prioritizes images with rich and distinctive visual features, ensuring robust feature matching and accurate geometric estimation [4]. Our empirical studies (Sec. IV-B) reveal that the criteria for these paradigms interact in both cooperative and competitive ways—what benefits one representation may hinder the performance of the other.

Game theory [19]–[21] provides a convenient language for describing trade-offs between decision-makers with competing objectives; prior robotics and SLAM works have used these ideas for planning and multi-agent estimation [22], [23]. We borrow the *intuition* of competing objectives from game theory to motivate our modelling of hybrid-representation keyframe selection: the discrete (feature-based) and field (radiance-field) subsystems prefer different frames. To avoid implying a formal multi-agent equilibrium, we treat this as a conceptual guide and implement an efficient, centralized solution — a dynamically weighted multi-objective scalarisation that selects keyframes by minimising a single time-varying weighted sum of representation-specific objectives.

III. Methodology

We emphasise from the outset that the proposed formulation is **inspired** by game-theoretic reasoning rather than claiming a

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

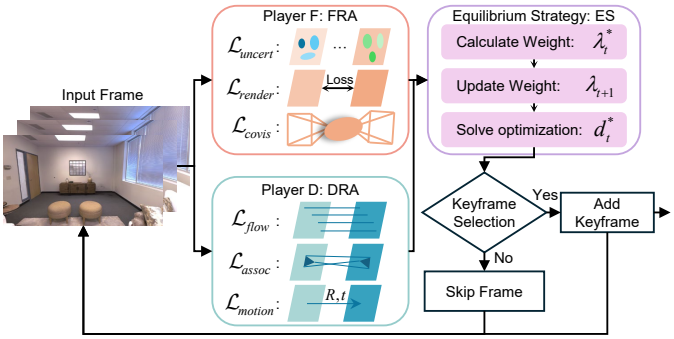

Fig. 2: Overview of Game-KFS. FRA and DRA compute evaluation signals from the renderer and the feature tracker respectively. A decision module forms the composite cost (Eq. 1) and outputs $d_t^\star$.

formal multi-agent equilibrium solution. Concretely, we model keyframe selection as an interaction between two conceptual *agents* — the **Field Representation Agent** (FRA) and the **Discrete Representation Agent** (DRA) — but implement a lightweight decision rule that combines the agents' evaluation signals via a dynamically-adapted scalarisation. This hybrid choice keeps the method practical for real-time VSLAM while retaining the intuition of competing/aligning objectives.

A. Overview

Fig. 2 summarises the pipeline. At each frame I_t both agents compute a small set of interpretable signals (costs) reflecting their own representation needs. The system forms two scalar objectives $\mathcal{A}_t(d_t)$ and $\mathcal{B}_t(d_t)$ (for FRA and DRA respectively) for the binary decision $d_t \in \{0,1\}$ (1 = select keyframe). A decision module then minimises a time-varying weighted sum of these objectives to produce the keyframe decision. The weight λ_t is adapted online from fast-to-compute signals so the scalarisation responds to scene and motion conditions.

B. Composite Decision

We operationalise the interaction via the following composite objective:

$$\mathcal{L}(d_t, \lambda_t) = \lambda_t \, \mathcal{A}_t(d_t) + (1 - \lambda_t) \, \mathcal{B}_t(d_t), \tag{1}$$

and select

$$d_t^\star = \arg\min_{d_t \in \{0,1\}} \mathcal{L}(d_t, \lambda_t). \tag{2}$$

As discussed in the manuscript, Eq. (1) is a time-varying weighted-sum scalarisation (a standard multi-objective technique). We use the term "game-theory–inspired" to stress that the two objectives represent distinct, sometimes competing, representation needs; the implemented decision is centralised for efficiency and real-time operation.

C. Field Representation Agent (FRA)

The FRA objective aggregates signals that measure how well the current frame helps the *radiance-field* style map improve rendering quality and reduce reconstruction uncertainty:

$$\mathcal{A}_t(d_t) = \beta_1 \, \mathcal{L}_{\text{uncert}} + \beta_2 \, \mathcal{L}_{\text{render}} + \beta_3 \, \mathcal{L}_{\text{covis}}. \tag{3}$$

a) Ray / pixel uncertainty $\mathcal{L}_{uncert}$.: We follow the uncertainty modelling used in recent field SLAM work (e.g., Uni-SLAM [24]). In a renderer that produces per-sample colour contributions along each ray (volume rendering / Gaussian splatting), let the (normalised) compositing weights along a ray be w_i and the per-sample predicted colours be c_i. The expected colour and its variance along the ray are:

$$\mathbb{E}[C] = \sum_i w_i c_i, \tag{4}$$

$$\mathrm{Var}[C] = \sum_i w_i \|c_i - \mathbb{E}[C]\|^2. \tag{5}$$

We convert the per-ray colour variance into a per-ray uncertainty score θ_m (e.g. by normalising by a fixed reference variance) and then form an image-level uncertainty:

$$\mathcal{L}_{\text{uncert}} = -\frac{1}{M} \sum_{m=1}^{M} \theta_m, \tag{6}$$

where M is the number of sampled rays (pixels). Intuitively, lower $\mathcal{L}_{\text{uncert}}$ indicates more confident rendering/prediction for that frame. The compositing weights w_i are computed in the standard NeRF / splatting fashion (transmittance T_i and alpha contributions), see e.g. NeRF literature for the full expressions; this yields a principled ray-level uncertainty that captures renderer epistemic/aleatoric spread.

b) Rendering fidelity $\mathcal{L}_{render}$.: Drawing inspiration from [25], we adopt a similarity metric to quantify the accuracy of the rendered scene. A high rendering quality score reflects a faithful reconstruction, while a low score indicates significant discrepancies. We quantify immediate rendering fidelity using a PSNR-based normalised term:

$$\mathcal{L}_{\text{render}} = 1 - \frac{\mathrm{PSNR}(I_t, \hat{I}_t)}{\mathrm{PSNR}_{\text{target}}}, \tag{7}$$

where $\hat{I}_t$ is the renderer output for the current estimated pose and $\mathrm{PSNR}_{\text{target}}$ is a normalising constant chosen so that $\mathcal{L}_{\text{render}}$ lies roughly in $[0,1]$ for typical scenes.

c) Gaussian co-visibility $\mathcal{L}_{covis}$.: Building upon the co-visibility formulation in [8], we incorporate a measure of scene overlap to assess the structural consistency between frames. A higher co-visibility score suggests greater structural consistency between the two frames. For Gaussian-splat style maps we measure overlap between the support of Gaussian primitives seen by the current frame and those associated to the most recent keyframe:

$$\mathcal{L}_{\text{covis}} = 1 - \frac{|V_t \cap V_{kf}|}{|V_t \cup V_{kf}|}, \tag{8}$$

where V_t and V_{kf} denote sets (or approximate supports) of Gaussians visible from each view. This term favours frames that increase coverage of under-observed scene regions.

D. Discrete Representation Agent (DRA)

The DRA objective captures traditional, feature-based tracking needs:

$$\mathcal{B}_t(d_t) = \alpha_1 \, \mathcal{L}_{\text{assoc}} + \alpha_2 \, \mathcal{L}_{\text{flow}} + \alpha_3 \, \mathcal{L}_{\text{motion}}. \tag{9}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

a) Data association $\mathcal{L}_{assoc}$.: To improve reproducibility we clarify the notation: n_{ref} are the reference-frame features (features in the current reference/keyframe), n_{match} are the successfully matched inliers between current frame and reference, n_{outlier} are rejected matches, and n_{total} is the number of detected features in the current frame. We use the compact score

$$\mathcal{L}_{\text{assoc}} = \frac{n_{\text{match}}}{n_{\text{ref}}} \exp\Big(- \frac{n_{\text{outlier}}}{n_{\text{total}}} \Big), \tag{10}$$

which rewards many correct matches while penalising high outlier ratios. (In our implementation reference features are taken from the most recent keyframe or from the tracker's active map depending on the front-end used, e.g., ORB-SLAM3-style bookkeeping [4].)

b) Optical-flow variation $\mathcal{L}_{flow}$.: Following the approach in [5], we assess motion smoothness by measuring the difference in optical flow vectors across frames. A lower variation suggests stable feature tracking, while a higher variation indicates potential motion inconsistencies or occlusions. We measure temporal consistency of tracked features:

$$\mathcal{L}_{\text{flow}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{u}_i - \mathbf{u}_{i-1}\|_2, \tag{11}$$

where $\mathbf{u}_i$ is the optical-flow vector of feature i at the current frame and $\mathbf{u}_{i-1}$ is the same feature's flow at the previous frame. Large variations indicate unstable tracking or occlusions.

c) Motion / displacement $\mathcal{L}_{motion}$.: Building upon the motion consistency assessment in [26], we incorporate both translational and rotational changes to ensure stable localization. A lower displacement change score suggests a more stable camera trajectory. We combine translational and rotational differences between the current pose and the reference/keyframe pose:

$$\mathcal{L}_{\text{motion}} = \|\mathbf{t}_t - \mathbf{t}_{kf}\|_2 + \omega \|\mathbf{R}_t - \mathbf{R}_{kf}\|_F, \tag{12}$$

where ω balances rotation vs. translation (we use a small ω when translation dominates the drift signal for our rigs).

E. Dynamic Weight Adaptation (Equilibrium Strategy)

Balancing $\mathcal{A}_t$ and $\mathcal{B}_t$ is handled by the adaptive weight λ_t. We form a candidate weight from fast-to-compute signals and apply exponential smoothing for temporal stability:

$$\lambda_t^\star = \sigma\big(\gamma_1 \, \mathcal{L}_{\text{assoc}}(I_t) + \gamma_2 \, \mathcal{L}_{\text{render}}(I_t)\big), \tag{13}$$

$$\lambda_{t+1} = \eta \, \lambda_t + (1 - \eta) \, \lambda_t^\star, \tag{14}$$

where $\sigma(\cdot)$ is a sigmoid that maps the candidate into $[0, 1]$, and η (we use $\eta = 0.8$) controls smoothing. Exponential moving average (EMA) is a standard, causal low-pass filter; it prevents rapid oscillations of λ_t while allowing timely adaptation to persistent changes in scene/motion conditions. EMA is commonly used in online perception systems because of its simplicity, stability and negligible computational cost.

We emphasise that $\mathcal{L}_{\text{render}}$ and $\mathcal{L}_{\text{uncert}}$ are *global* in the sense that they are computed from the renderer's outputs and therefore reflect the global map state; this is how our update indirectly incorporates global information (e.g., changing rendering uncertainty after a map update or after some background optimisation). In practice the renderer runs asynchronously for heavy optimisation while lightweight render checks used for $\mathcal{L}_{\text{render}}$ and $\mathcal{L}_{\text{uncert}}$ are performed at reduced resolution or frequency so as not to block the main loop.

F. Stability and Interpretation

Because the decision is centralised, the chosen $d_t^\star$ is *stable* under the current λ_t in the sense that switching the single binary choice would not reduce $\mathcal{L}$ (Eq. 1). We therefore avoid claiming a formal decentralised Nash equilibrium; instead, "Nash-like" or "stable decision under λ_t" is used as an explanatory shorthand. The relation between centralised scalarisation and multi-agent equilibria is well studied in the optimisation and game-theory literature (see e.g. [27]).

IV. Experimental Evaluation

To validate the effectiveness of the proposed keyframe selection strategy in hybrid representation VSLAM, we conduct experiments on multiple datasets and in real-world robotic environments. We first introduce the experimental setup and implementation details. Then, we perform experiments to validate the motivation behind our approach. Next, we conduct comparative experiments on two public datasets. Subsequently, we evaluate the method's performance on a real robot platform. Finally, we carry out ablation studies and sensitivity analyses.

A. Experimental Setup and Implementation

1) Datasets: We evaluate the proposed method on multiple sequences from two widely used public datasets, Replica [28] and TUM [29]. Additionally, we collect real-world scene data using our laboratory's mobile robot platform, as shown in Fig. 3.

2) Baselines: We compare our proposed method, *Game-KFS*, with three baseline methods: *SplaTAM* [9], which selects a fixed number of keyframes at regular intervals; *Photo-SLAM* [13], which focuses on optimizing discrete representation; and *MonoGS* [8], which aims to optimize field representation. In contrast, *Game-KFS* adaptively balances the characteristics of both discrete and field representations for keyframe selection.

3) Implementation Details: Our primary framework is implemented based on *Photo-SLAM* [13]. Similar to [11], after the 3D Gaussian Splatting (3D GS) model is optimized, we use this model to refine the pose. To ensure that performance is not affected by different systems, in addition to the experiments in Sec. IV-B, we re-implement the keyframe selection methods of *SplaTAM* [9], *Photo-SLAM* [13], and *MonoGS* [8] within the *Photo-SLAM* [13] framework. All experiments are conducted on a server equipped with an Intel® Xeon(R) Silver 4214R CPU and an NVIDIA GeForce RTX 4090 GPU. For the *SplaTAM* [9], we set the keyframe sampling interval to one frame every 10 frames. We set $\beta_1 = 0.3$, $\beta_2 = 0.3$, $\beta_3 = 0.4$, $\alpha_1 = 0.5$, $\alpha_2 = 0.3$, and $\alpha_3 = 0.2$.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

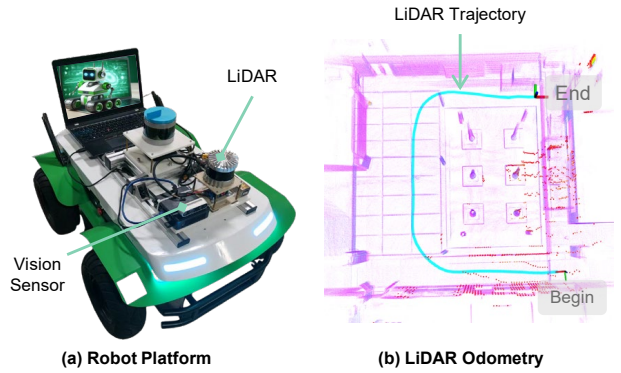

Fig. 3: Real robot platform and LiDAR odometry ground truth. (a) The physical robot platform equipped with a LiDAR and a vision sensor. (b) The LiDAR odometry trajectory, which serves as the ground truth for evaluating visual localization.

TABLE I: Comparison of keyframe selection methods in MonoGS.

Sequence	Method	Total Images	Keyframe Num	ATE RMSE (cm) ↓	PSNR (dB) ↑
fr1/desk	MonoGS+*DRD*	613	97	75.97	17.09
	MonoGS+*FRD*	613	128	**4.02**	**22.05**
fr3/office	MonoGS+*DRD*	2585	292	4.75	18.81
	MonoGS+*FRD*	2585	246	**4.13**	**21.93**

TABLE II: Comparison of keyframe selection in ORB-SLAM3. The symbol "-" denotes data being insignificant when below 80% [3].

Sequence	Method	Total Images	Keyframe Num	Tracking Rate ↑	ATE RMSE (cm) ↓
fr1/desk	ORB-SLAM3+*DRD*	613	108	**96.90%**	**1.75**
	ORB-SLAM3+*FRD*	613	82	72.53%	-
fr3/office	ORB-SLAM3+*DRD*	2585	286	99.65%	**1.54**
	ORB-SLAM3+*FRD*	2585	231	**99.69%**	115.55

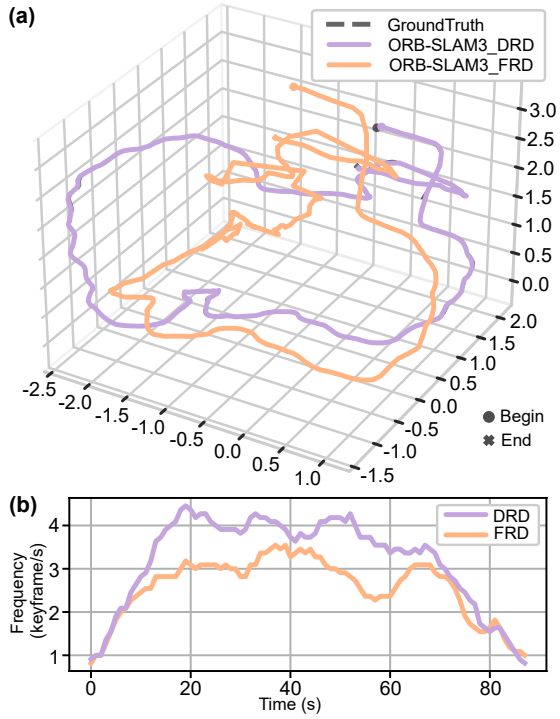

Fig. 4: Keyframes trajectory and frequency comparison. The experiment is conducted on the TUM fr3/office sequence. (a) Trajectory comparison across different methods. (b) Comparison of keyframe frequency over time for two different selection methods. The curves has been smoothed.

B. Verification of Preliminary Hypothesis

To verify our hypothesis that keyframe selection methods driven by a single type of representation do not perform well in variant representation systems, we conduct both quantitative and qualitative evaluations. First, we extract keyframes using a discrete representation-driven keyframe selection method (*DRD*) within *ORB-SLAM3* [4]. We then apply these selected keyframes to a field representation-based VSLAM system (*MonoGS* [8]). The evaluation is performed on the *fr1/desk* and *fr3/office* sequences of the TUM dataset, with results presented in Table I. The results indicate that the *DRD* performs poorly when applied to a field representation-based VSLAM system. Specifically, on the *fr1/desk* sequence, the ATE RMSE reaches 75.97 cm, which is approximately 19 times higher than the error of the field representation-driven keyframe selection method. Additionally, the PSNR is only 17.09 dB, while the SSIM and LPIPS are 0.58 and 0.51, respectively.

Next, we perform a reciprocal experiment, where keyframes are first selected using the field representation-driven keyframe selection method (*FRD*) within *MonoGS*. These keyframes are then applied to the discrete representation-based VSLAM system (*ORB-SLAM3*). The evaluation is conducted on the same *fr1/desk* and *fr3/office* sequences, and the results are summarized in Table II and Fig. 4. As shown in Table II, *FRD* exhibits poor performance in a discrete representation-based VSLAM system. On the *fr1/desk* sequence, the tracking success rate is only 72.53%, which is classified as a tracking failure [3]. On the *fr3/office* sequence, the ATE RMSE reaches 115.55 cm, further demonstrating the incompatibility of *FRD* in this setting. Fig. 4(a) presents the trajectory and keyframe sequence visualization, where it is evident that the keyframe selection of the *FRD* is suboptimal, leading to significant errors. Fig. 4(b) demonstrates that the *DRD* and *FRD* keyframe selection methods exhibit a pronounced selective bias towards different regions of the image sequence.

The results from both experiments demonstrate that *DRD* does not perform well in field representation-based VSLAM, and conversely, *FRD* is ineffective in discrete representation-based VSLAM. This highlights the limitation of single-representation-driven keyframe selection methods and underscores the necessity of developing a novel keyframe selection strategy that can adapt to different representation systems. Such an approach is crucial in hybrid representation VSLAM, as selecting one keyframe impacts both discrete and field representations.

C. Evaluation of Keyframe Selection Strategy on Replica

We begin our evaluation of keyframe selection strategies on a high-fidelity synthetic dataset, Replica [28], which features multiple indoor scenes. In these experiments, we compare different keyframe selection methods across six dataset sequences, and the results are summarized in Table III. Our

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE III: Quantitative results on the Replica dataset. We mark the best two results with **first** and second.

On Replica		SplaTAM	Photo-SLAM	MonoGS	**Ours**
Office0	Keyframe Num	158	132	346	142
	RMSE (cm) ↓	0.48	0.44	44.05	**0.35**
	PSNR ↑	38.88	32.49	34.92	**41.21**
	SSIM ↑	0.96	0.91	0.94	**0.98**
	LPIPS ↓	0.045	0.063	0.15	**0.033**
Office1	Keyframe Num	121	110	272	122
	RMSE (cm) ↓	0.31	0.31	16.19	**0.22**
	PSNR ↑	39.28	34.18	34.72	**41.68**
	SSIM ↑	0.96	0.93	0.93	**0.97**
	LPIPS ↓	0.051	0.057	0.17	**0.031**
Office2	Keyframe Num	201	171	393	165
	RMSE (cm) ↓	1.79	4.06	43.11	**0.95**
	PSNR ↑	32.99	32.99	26.62	**35.65**
	SSIM ↑	0.94	0.94	0.88	**0.99**
	LPIPS ↓	0.079	0.083	0.24	**0.043**
Room0	Keyframe Num	167	132	360	155
	RMSE (cm) ↓	0.35	**0.33**	11.37	0.34
	PSNR ↑	31.53	30.18	27.29	**32.85**
	SSIM ↑	0.91	0.89	0.84	**0.95**
	LPIPS ↓	**0.073**	0.088	0.19	0.078
Room1	Keyframe Num	177	175	331	168
	RMSE (cm) ↓	**1.61**	2.48	36.11	1.87
	PSNR ↑	33.48	31.33	27.49	**35.65**
	SSIM ↑	0.92	0.91	0.83	**0.94**
	LPIPS ↓	0.058	0.089	0.27	**0.051**
Room2	Keyframe Num	196	153	411	162
	RMSE (cm) ↓	0.26	0.22	4.57	**0.18**
	PSNR ↑	34.71	34.05	35.18	**37.86**
	SSIM ↑	0.95	0.94	0.95	**0.97**
	LPIPS ↓	0.054	0.063	0.11	**0.038**

Game-KFS method achieves superior performance in most scenarios, demonstrating lower ATE RMSE, higher PSNR and SSIM, and lower LPIPS compared to the baseline methods. The improvement of localization accuracy mainly relies on the fact that after we obtain better rendering results, we then use these better results to optimize the pose. Additionally, *Game-KFS* selects keyframes more efficiently by identifying a smaller yet more effective set of keyframes. For instance, in the *Office0* sequence, our method selects only 142 keyframes, whereas *Photo-SLAM* selects 346 keyframes. This highlights the efficiency of our approach in selecting representative keyframes.

It is noteworthy that since hybrid representation VSLAM primarily relies on discrete representation-based tracking, the *MonoGS* method, which prioritizes field representation-based keyframe selection, exhibits suboptimal performance across most sequences. Meanwhile, the *SplaTAM* method does not inherently favor any specific representation, leading to better overall performance than both *Photo-SLAM* and *MonoGS* in most sequences. This further underscores the limitations of single representation-driven keyframe selection methods in hybrid representation VSLAM systems.

TABLE IV: Quantitative results on the TUM dataset.

On TUM		SplaTAM	Photo-SLAM	MonoGS	**Ours**
fr1/desk	Keyframe Num	74	80	128	82
	RMSE (cm) ↓	2.23	**1.69**	3.61	1.72
	PSNR ↑	20.29	19.58	20.77	**23.12**
	SSIM ↑	0.73	0.71	0.70	**0.89**
	LPIPS ↓	0.26	0.29	0.35	**0.21**
fr2/xyz	Keyframe Num	51	51	112	70
	RMSE (cm) ↓	0.74	0.76	4.21	**0.65**
	PSNR ↑	21.35	22.34	22.21	**25.12**
	SSIM ↑	0.73	0.72	0.73	**0.94**
	LPIPS ↓	0.15	0.19	0.28	**0.12**
fr3/office	Keyframe Num	268	349	247	220
	RMSE (cm) ↓	**1.16**	86.99	3.89	1.25
	PSNR ↑	19.68	22.47	17.34	**25.31**
	SSIM ↑	0.69	0.77	0.66	**0.86**
	LPIPS ↓	0.24	0.42	0.34	**0.15**

D. Evaluation of Keyframe Selection Strategy on TUM

In addition, we examine the performance of various keyframe selection strategies in challenging real scenarios using the TUM dataset. Specifically, the *fr1/desk*, *fr2/xyz*, and *fr3/office* sequences are employed to assess these methods, with quantitative outcomes presented in Table IV. As shown in Table IV, our *Game-KFS* method consistently outperforms baseline approaches in most scenarios, achieving lower ATE RMSE, higher PSNR and SSIM, and lower LPIPS. Furthermore, our method selects keyframes more efficiently by identifying a more compact yet representative set of keyframes. For example, in the *fr3/office* sequence, *Game-KFS* selects only 220 keyframes, whereas *Photo-SLAM* selects 349 keyframes. This result highlights the efficiency of our approach in maintaining accurate scene representation with fewer keyframes.

Similar to the findings in Table III, the *MonoGS* method performs suboptimally across most sequences, reinforcing its limitations in hybrid representation VSLAM. Meanwhile, the *SplaTAM* method achieves better overall performance than both *Photo-SLAM* and *MonoGS*, further demonstrating the drawbacks of single-representation-driven keyframe selection methods in hybrid representation VSLAM systems.

E. Evaluation of Performance on Real Robot

Furthermore, to validate the effectiveness of the proposed method in practical robotic applications, we perform experiments on a mobile robot platform in our laboratory. As shown in Fig. 3, the platform is equipped with both a vision sensor and a LiDAR module, where LiDAR-based odometry [30] serves as a reference for evaluating the localization accuracy of the VSLAM system.

The quantitative tracking and mapping performance results are summarized in Table V. Our method demonstrates superior performance, particularly in rendering speed, achieving the capability to render thousands of frames per second. As shown in Fig. 5, we plot the trajectories of the selected keyframes. Compared to a fixed-interval sampling strategy (the *SplaTAM* keyframe-selection policy, re-implemented within the *Photo-SLAM* framework), our method adaptively selects keyframes

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE V: Quantitative results on the real robot.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Rendering FPS ↑	RMSE (m) ↓
Ours	**23.96**	**0.77**	**0.23**	1014.32	**0.31**
Photo-SLAM	21.84	0.73	0.28	**1134.12**	0.48

TABLE VI: Running time on the real robot. All times are reported in milliseconds (ms).

Method	Tracking	Local Mapping	Rendering	Keyframe Selection	Main-loop
Ours	22.95	225.75	0.98	2.12	32.78
Photo-SLAM	20.89	200.95	0.89	0.05	30.65

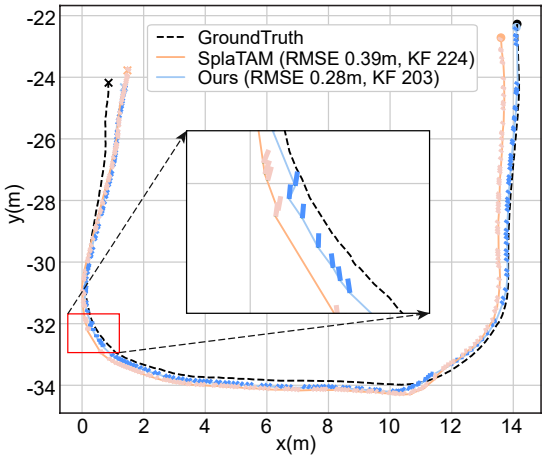

Fig. 5: Keyframe trajectory comparison between our method and fixed-interval keyframe selection (*SplaTAM*). Each orange rectangle or blue rectangle represents an individual keyframe.

suited to the scene, for example at the turns highlighted by the red rectangle. In addition, Fig. 6 presents the qualitative mapping results obtained using our approach, showcasing its ability to generate visually realistic reconstructions.

Table VI reports per-module running times measured on the real robot (all values in ms). Our pipeline records Tracking = 22.95 ms, Local Mapping = 225.75 ms, Rendering = 0.98 ms, Keyframe Selection (Game-KFS) = 2.12 ms, yielding an effective main-loop time of ≈32.78 ms. The Photo-SLAM baseline measures Tracking = 20.89 ms, Local Mapping = 200.95 ms, Rendering = 0.89 ms, Keyframe Selection = 0.05 ms, and a main-loop time of ≈30.65 ms. Thus, the dedicated Game-KFS decision logic introduces a modest per-frame overhead (≈+2.07 ms compared to Photo-SLAM's selection), but the resulting main-loop time for our full pipeline (≈32.8 ms) remains within the 30 Hz sensor budget (≈33.3 ms/frame) on the tested hardware. Note that Local Mapping is executed asynchronously and is reported here for completeness; its background cost does not block the main tracking loop.

F. Ablation Study and Sensitivity Analyses

The ablation results, presented in Table VII, demonstrate the importance of each module in improving system performance. The sensitivity analyses results are shown in Table VIII.

(a) Ground Truth (b) Ours (c) Photo-SLAM

Fig. 6: Qualitative mapping results on the real robot.

TABLE VII: Ablation study. The RMSE is measured in cm.

On Replica			Office3			Office4		
DRA	FRA	ES	RMSE ↓	PSNR ↑	LPIPS ↓	RMSE ↓	PSNR ↑	LPIPS ↓
✓	✓	✓	**0.31**	**37.65**	0.021	**0.42**	**36.87**	0.074
✗	✓	✓	0.42	34.21	**0.015**	0.54	34.65	0.12
✓	✗	✓	0.75	32.45	0.054	0.82	29.65	0.087
✓	✓	✗	0.65	36.25	0.15	0.68	28.65	**0.065**

TABLE VIII: Sensitivity Analysis of α and β on TUM-fr1/desk and Replica-Office0. Reported metrics are mean±std over 3 runs.

Sequence	Setting (weights)	ATE RMSE (cm)	PSNR (dB)
TUM-fr1/desk	$\alpha = (0.20, 0.48, 0.32)$	1.82 ± 0.04	22.95 ± 0.07
	$\alpha = (0.50, 0.30, 0.20)$	1.71 ± 0.05	23.15 ± 0.08
	$\alpha = (0.80, 0.12, 0.08)$	1.92 ± 0.06	22.85 ± 0.09
	$\beta = (0.40, 0.40, 0.20)$	1.80 ± 0.05	23.00 ± 0.06
	$\beta = (0.30, 0.30, 0.40)$	1.71 ± 0.05	23.15 ± 0.08
	$\beta = (0.20, 0.20, 0.60)$	1.92 ± 0.06	22.85 ± 0.09
	$\beta = (0.10, 0.10, 0.80)$	1.95 ± 0.08	21.15 ± 0.12
Replica-Office0	$\alpha = (0.20, 0.48, 0.32)$	0.36 ± 0.02	37.85 ± 0.35
	$\alpha = (0.50, 0.30, 0.20)$	0.34 ± 0.02	40.95 ± 0.21
	$\alpha = (0.80, 0.12, 0.08)$	0.38 ± 0.02	39.95 ± 0.12
	$\beta = (0.40, 0.40, 0.20)$	0.35 ± 0.02	36.68 ± 0.35
	$\beta = (0.30, 0.30, 0.40)$	0.34 ± 0.02	40.95 ± 0.21
	$\beta = (0.20, 0.20, 0.60)$	0.38 ± 0.02	38.76 ± 0.12
	$\beta = (0.10, 0.10, 0.80)$	0.41 ± 0.04	36.15 ± 0.15

1) Impact of DRA: We first examine the role of the *DRA* module, which enhances keyframe selection from the perspective of discrete representations. As shown in Table VII, removing DRA (✗, ✓, ✓) leads to a significant performance drop in *Office3*, with RMSE increasing from **0.31** to 0.42 and PSNR decreasing from **37.65** to 34.21. Similarly, in *Office4*, RMSE rises from **0.42** to 0.54, and LPIPS increases from **0.074** to 0.12. These results highlight the effectiveness of DRA in improving localization accuracy and maintaining high-quality reconstruction.

2) Impact of FRA: Next, we evaluate *FRA*, which optimizes keyframe selection from the perspective of field representations. When *FRA* is disabled (✓, ✗, ✓), system performance deteriorates substantially. In *Office3*, RMSE increases from **0.31** to 0.75, PSNR drops from **37.65** to 32.45, and LPIPS rises from **0.021** to 0.054. Similarly, in *Office4*, RMSE increases to 0.82, and PSNR declines to 29.65. These results indicate that *FRA* plays a crucial role in improving the perceptual quality of the reconstructed scene.

3) Impact of ES: Finally, we investigate the effect of *ES*, which dynamically adjusts the threshold for keyframe selec-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

tion. Disabling *ES* (✓, ✓, ✗) leads to noticeable performance degradation, especially in LPIPS and PSNR. In *Office3*, LPIPS increases from **0.021** to 0.15, while in *Office4*, PSNR decreases from **36.87** to 28.65, suggesting that *ES* enhances adaptability to different scene conditions.

4) Sensitivity Analysis: Table VIII summarises the sensitivity of the system to different weight configurations. Across both TUM-fr1/desk and Replica-Office0, the results show that the default settings $\alpha = (0.50, 0.30, 0.20)$ and $\beta = (0.30, 0.30, 0.40)$ provide the best balance between localisation accuracy and rendering quality. Moderate perturbations around these defaults lead to only small changes in ATE and PSNR, indicating robustness. In contrast, extreme biasing of a single component (e.g., $\alpha_1 = 0.80$ or $\beta_3 = 0.80$) degrades performance by up to 10–20%.

V. Conclusion

In this paper, we propose a keyframe selection method inspired by game theory for hybrid representation VSLAM. By adapting to the distinct characteristics of discrete and field representations, our method enables efficient keyframe selection that improves both localization accuracy and scene rendering quality. Experimental results on multiple public datasets demonstrate superior performance across diverse scenarios, and validation on a real robotic platform confirms the method's practicality. We will further extend the application of *Game-KFS*, including its integration into edge-cloud collaborative hybrid representation VSLAM [31].

We note that the proposed approach is game-theory-inspired rather than a derivation from a formal game-theoretic model; developing a rigorous game-theoretic formulation and accompanying theoretical analysis remains an interesting direction.

References

[1] W. Chen, S. Chen, J. Leng, J. Wang, Y. Guan, M. Q.-H. Meng, and H. Zhang, "A review of cloud-edge SLAM: Toward asynchronous collaboration and implicit representation transmission," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 15 437–15 453, Nov. 2024.

[2] S. Chen, X. Luo, Z. Lin, S. Wen, H. Zhang, and W. Chen, "Bridging the gap between explicit and implicit representations: Cross-data association for VSLAM," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 21 252–21 266, Dec. 2024.

[3] W. Chen, Z. Lin, L. Zhu, S. Chen, Y. Guan, and H. Zhang, "Cloud-edge collaborative submap-based VSLAM using implicit representation transmission," *IEEE Trans. Veh. Technol.*, vol. 73, no. 10, pp. 14 537–14 546, Oct. 2024.

[4] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[5] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[6] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "NICE-SLAM: Neural implicit scalable encoding for SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12 776–12 786.

[7] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, Jul. 2023.

[8] H. Matsuki, R. Murai, P. H. J. Kelly, and A. J. Davison, "Gaussian splatting SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18 039–18 048.

[9] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, "SplaTAM: Splat, track & map 3D Gaussians for dense RGB-D SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 21 357–21 366.

[10] S. Chen, S. Ji, L. Zhu, X. Zhou, H. Zhang, W. Chen, and Y. Guan, "X-RepSLAM: VLM-driven adaptive cross-representation visual SLAM," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 20 000–20 018, 2025.

[11] C.-M. Chung, Y.-C. Tseng, Y.-C. Hsu, X. Q. Shi, Y.-H. Hua, J.-F. Yeh, W.-C. Chen, Y.-T. Chen, and W. H. Hsu, "Orbeez-SLAM: A real-time monocular visual SLAM with ORB features and NeRF-realized mapping," in *Proc. Int. Conf. Robot. Autom.*, 2023, pp. 9400–9406.

[12] Y. Mao, X. Yu, K. Wang, Y. Wang, R. Xiong, and Y. Liao, "NGEL-SLAM: Neural implicit representation-based global consistent low-latency SLAM system," in *Proc. Int. Conf. Robot. Autom.*, 2024, pp. 6952–6958.

[13] H. Huang, L. Li, H. Cheng, and S.-K. Yeung, "Photo-SLAM: Real-time simultaneous localization and photorealistic mapping for monocular, stereo, and RGB-D cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 21 584–21 593.

[14] J. Naumann, B. Xu, S. Leutenegger, and X. Zuo, "NeRF-VO: Real-time sparse visual odometry with neural radiance fields," *IEEE Robot. Autom. Lett.*, vol. 9, no. 8, pp. 7278–7285, Aug. 2024.

[15] P. Zhu, Y. Zhuang, B. Chen, L. Li, C. Wu, and Z. Liu, "MGS-SLAM: Monocular sparse tracking and Gaussian mapping with depth smooth regularization," *IEEE Robot. Autom. Lett.*, vol. 9, no. 11, pp. 9486–9493, Nov. 2024.

[16] H. Zhou, Z. Guo, Y. Ren, S. Liu, L. Zhang, K. Zhang, and M. Li, "MoD-SLAM: Monocular dense mapping for unbounded 3D scene reconstruction," *IEEE Robot. Autom. Lett.*, vol. 10, no. 1, pp. 484–491, Jan. 2025.

[17] W. Chen, H. Ye, L. Zhu, C. Tang, C. Fu, Y. Chen, and H. Zhang, "Keyframe selection with information occupancy grid model for long-term data association," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 2786–2793.

[18] B. Mildenhall et al., "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Dec. 2022.

[19] M. Wang, N. Mehr, A. Gaidon, and M. Schwager, "Game-theoretic planning for risk-aware interactive agents," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 6998–7005.

[20] Y. Cao, X. Zeng, and Z. Yin, "A game theoretic decision-making framework with conflict-aware Nash equilibrium selection for autonomous vehicles at uncontrolled intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 1, pp. 210–224, Jan. 2025.

[21] M. Mohammadi, R. Tavakkoli-Moghaddam, A. Siadat, and Y. Rahimi, "A game-based meta-heuristic for a fuzzy bi-objective reliable hub location problem," *Eng. Appl. Artif. Intell.*, vol. 50, pp. 1–19, Apr. 2016.

[22] R. Chandra, M. Wang, M. Schwager, and D. Manocha, "Game-theoretic planning for autonomous driving among risk-aware human drivers," in *Proc. Int. Conf. Robot. Autom.*, 2022, pp. 2876–2883.

[23] C.-Y. Chiu and D. Fridovich-Keil, "GTP-SLAM: Game-theoretic priors for simultaneous localization and mapping in multi-agent scenarios," in *Proc. IEEE Conf. Decis. Control*, 2022, pp. 247–252.

[24] S. Wang, Y. Xie, C.-P. Chang, C. Millerdurai, A. Pagani, and D. Stricker, "Uni-SLAM: Uncertainty-aware neural implicit SLAM for real-time dense indoor scene reconstruction," in *Proc. IEEE/CVF Winter Conf. Appli. Comput. Vis.*, 2025, pp. 2228–2239.

[25] J. Wilkinson, J. Naylor, R. Griffiths, and D. G. Dansereau, "Adaptive keyframe selection for online iterative NeRF construction," in *Proc. RoboNerF: 1st Workshop On Neural Fields In Robotics at ICRA*, 2024, pp. 1–5.

[26] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, Apr. 2017.

[27] K. Miettinen, *Nonlinear multiobjective optimization*. Springer Science & Business Media, 1999, vol. 12.

[28] J. Straub, et al., "The Replica dataset: A digital Replica of indoor spaces," Jun. 2019. [Online]. Available: https://arxiv.org/abs/1906.05797

[29] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 573–580.

[30] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "FAST-LIO2: Fast direct LiDAR-inertial odometry," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2053–2073, 2022.

[31] W. Chen, X. Luo, X. Huo, S. Chen, J. Li, C. L. P. Chen, and H. Zhang, "VC-SLAM: Optimizing cloud-edge VSLAM transmission based on variable-order Chebyshev-KAN," *IEEE/ASME Trans. Mechatron.*, vol. Early Access, pp. 1–11, 2025.